

DOCUMENT RESUME

ED 186 456

TM 800 145

AUTHOR Morgan, George
TITLE A Criterion-Referenced Measurement Model with Corrections for Guessing and Carelessness. Occasional Paper No. 13.
INSTITUTION Australian Council for Educational Research, Hawthorn.
PUB DATE Jun 79
NOTE 88p.; Best copy available. Appendices D to G are available as a separate document from the same source.
AVAILABLE FROM Australian Council for Educational Research Limited, P.O. Box 210, Hawthorn, Victoria, Australia 3122 (\$5.00)
EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Basic Skills; *Bayesian Statistics; Computer Programs; *Criterion Referenced Tests; *Cutting Scores; Elementary Secondary Education; Foreign Countries; *Guessing (Tests); *Latent Trait Theory; *Mathematical Models; National Surveys; Quantitative Tests; Verbal Tests
IDENTIFIERS Australia; *Carelessness

ABSTRACT

This study developed a two-parameter linear probabilistic model of criterion referenced measurement, which relates an examinee's observed ability to his true ability through a guessing parameter and a carelessness parameter, both of which are probabilities. This model was applied to a survey of 10- and 14-year-old Australian students to estimate the proportion who had been able to master basic skills in mathematics and literacy. The estimates were derived by using a Bayesian inferential procedure incorporating the developed measurement model within a beta-binomial, decision-theoretic framework based on a two-state conception of mastery and non-mastery. The results of a computer simulation of the model indicate that guessing and carelessness may markedly affect the determination of passing scores, and hence the accuracy of decision about mastery. The computer program for calculating cutting scores, posterior probabilities, and prior and posterior distribution statistics is appended. (Author/CTM)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

TM800 145

ED186456

A Criterion-referenced Measurement Model with Corrections for Guessing and Carelessness

George Morgan

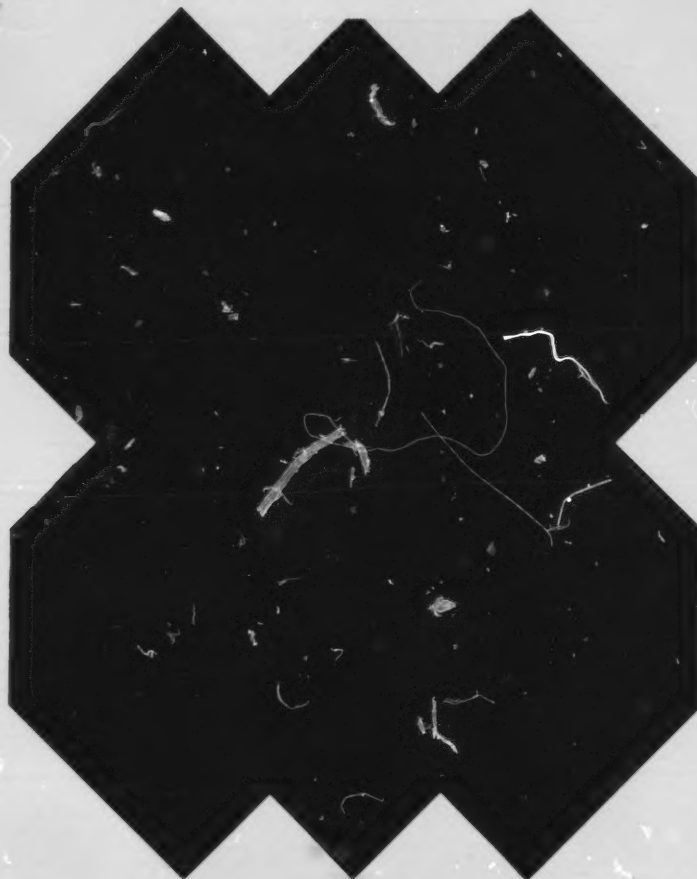
U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

"PERMISSION TO REPRODUCE THIS
MATERIAL IN MICROFICHE ONLY
HAS BEEN GRANTED BY

J. P. Keeves

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC) AND
USERS OF THE ERIC SYSTEM"

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY



Occasional Paper No.13 - June 1979

A CRITERION-REFERENCED MEASUREMENT MODEL
WITH CORRECTIONS FOR GUESSING
AND CARELESSNESS

BEST COPY AVAILABLE

A CRITERION-REFERENCED MEASUREMENT MODEL
WITH CORRECTIONS FOR GUESSING
AND CARELESSNESS

by

George Morgan

Occasional Paper No.13 - June 1979

The Australian Council for Educational Research Limited
Frederick Street, Hawthorn 3122 Victoria

Published by The Australian Council for Educational Research Limited
Frederick Street, Hawthorn, Victoria, 3122.
Printed and bound by
Jenkin Buxton Printers Pty Ltd, 1-13 Abbotsford Street, West Melbourne, 3003.

Cataloguing-in-Publication data.

Morgan, George

A criterion-referenced measurement model with corrections for guessing
and carelessness.

(Occasional paper; No.13)
ISBN 0 85563 192 9

1. Educational tests and measurement I. Title. (Series; Australian
Council for Educational Research. Occasional paper; No.13)

371.26

Copyright © ACER 1979

No part of this book may be reproduced in any form without permission
from the publisher.

PREFACE

This study was carried out as part of the program of the Measurement and Evaluation Division, and was part of the Core Program supported by the annual grants received by the ACER from the State Governments of the Australian Commonwealth and the Australian Government. Its main goal was the development of a criterion-referenced measurement model, within a Bayesian decision-theoretic framework, which could help with the interpretation of criterion-referenced test data by taking into account systematic errors due to guessing and carelessness.

In order to reduce the length of this publication only Appendices A, B and C from the original report have been reproduced here. Appendices D to G referred to in this report are available as a separate document entitled: A Criterion-Referenced Measurement Model with Corrections for Guessing and Carelessness - Supplementary Appendices. This document may be obtained from the ACER, the price being \$2.00 including postage.

TABLE OF CONTENTS

	Page
LIST OF TABLES	viii
LIST OF FIGURES	ix
ACKNOWLEDGMENTS	xi
CHAPTER 1 INTRODUCTION	1
Concepts of Criterion-Referenced Measurement and Tests	3
Measurement of Mastery	7
Bayesian Formulation of Criterion-Referenced Measurement	8
Extraneous Variables in Criterion-Referenced Measurement	11
Origin and Purpose of the Present Study	13
CHAPTER 2 A MODEL OF CRITERION-REFERENCED MEASUREMENT WITH CORRECTIONS FOR GUESSING AND CARELESSNESS	15
Binary Response Test Data	16
The Model: Relation Between Observed Ability and True Ability	17
CHAPTER 3 A BAYESIAN DECISION-THEORETIC PROCEDURE FOR ESTIMATING MASTERY STATUS OF AN EXAMINEE	25
The Two-Action Decision Problem	25
Bayesian Solution of the Two-Action Decision Problem	27
The Beta-Binomial Model	30
CHAPTER 4 TRIAL CALCULATIONS: A COMPUTER SIMULATION OF THE MODEL	34
Method	34
Parameters of Prior Distributions	36
Posterior Probabilities	37
Cutting Scores	38
Statistics of Posterior Distributions	39
CHAPTER 5 PRACTICAL APPLICATION OF THE MODEL: THE DETERMINATION OF CUTTING SCORES FOR CRITERION-REFERENCED TESTS FROM THE AUSTRALIAN STUDIES IN SCHOOL PERFORMANCE	40
Tests and Sub-Tests of Literacy and Numeracy	40
Specifying Prior Distributions for Tests and Sub-Tests	42
Cutting Scores for Tests and Sub-Tests	46

CHAPTER 6	CONCLUSION	Page 50
APPENDIX A1	Distribution of Scores on Reading Sub-Tests and Whole Test for 10-Year-Old Australian Students	54
APPENDIX A2	Distribution of Scores on Reading Sub-Tests and Whole Test for 14-Year-Old Australian Students	55
APPENDIX A3	Distribution of Scores on Numeration Sub-Tests and Whole Test for 10-Year-Old Australian Students	56
APPENDIX A4	Distribution of Scores on Numeration Sub-Tests and Whole Test for 14-Year-Old Australian Students	57
APPENDIX B	Sub-Tests of the Whole Tests	58
APPENDIX C	Program BAYES for Calculating: (a) Cutting Scores; (b) Posterior Probabilities $\Pr(\theta_0 \geq \theta_c \mid x = c)$; and (c) Means, Modes, and Standard Deviations of Prior Distributions $p(\theta_0)$, and Posterior Distirubitons $f(\theta_0 \mid x = c)$	60
BIBLIOGRAPHY		72

LIST OF TABLES

		Page
Table 3.1	Losses $L_t(d_i, \theta_t)$ Associated with Incorrect Actions (θ_t metric)	26
Table 3.2	Losses $L_o(d_i, \theta_o)$ Associated with Incorrect Actions (θ_o metric)	26
Table 4.1	θ_o Intervals for Stated Values of β and γ	36
Table 5.1	Posterior Probability that Examinee's Observed Ability is Greater than $\bar{\theta}_c$ given a Uniform Prior Distribution	43
Table 5.2	Selected Informative Beta Prior Distributions for Tests and Sub-Tests	45
Table 5.3	Estimated Cutting Scores for Tests Based on Three Different Prior Distributions	47
Table 5.4	Cutting Scores (c) for the Reading and Numeration Sub-Tests and Whole Tests	49

LIST OF FIGURES

	Page
Figure 2.1 Plots of the Relation $\theta_0 = \beta + \theta_t(1 - \beta)(1 - \gamma)$ for Stated Values of γ and $\beta = 0.25$	23
Figure 4.1 Examples of Graphs of $\rho(\theta_t) = B(a,b)$ where $B(a,b)$ is a Beta Function with Parameters a and b	35

ACKNOWLEDGMENTS

The details of the Criterion-referenced model developed in this study were a direct outcome of discussions with Dr H. Lindman of Indiana University, Indiana, USA, and Dr D.F. Burrill of the Ontario Institute for Studies in Education, Canada. To them I extend many thanks for their useful suggestions. In addition, I wish to express my appreciation to Dr Terry Dunn, of the Centre for the Study of Higher Education, for his guidance and help.

I would also like to thank Mr S.F. Bourke of ACER, for the many discussions I have had with him concerning various facets of this study, and for his critical reading of this report.

G. Morgan

CHAPTER 1

INTRODUCTION

Criterion-referenced testing is currently finding much application in assessment programs which attempt to measure validly and reliably what the individual knows and what the individual is capable of doing, perhaps after completing a unit of instruction in a certain subject. It is a procedure which is seen to be useful for obtaining the kinds of measures of student achievement and their acquisition of skills which facilitates decision making in objectives-based instructional programs - programs that are designed to maximize the efficiency of the teaching-learning process, especially at the level of the individual learner.

By using a criterion-referenced test, the examiner wishes to obtain information about the individual's standing with respect to a well-defined body of knowledge and behaviours. This standing is often represented in terms of a measure of performance (competency, or functioning) based on the individual's test data, which is often, but not always, independent of the performances of others that might be tested together with that individual. Such information can be used to make a variety of decisions, individual and programmatic, such as the placement of individuals within a program of learning, the revision of learning materials or the rearrangement of learning units. For example, in self-paced instructional programs, which are usually objectives-based, students whose levels of performance are satisfactory, as measured by a criterion-referenced test, will be promoted to the next unit of the instructional sequence, whereas students whose levels of performance are unsatisfactory will be retained on the current unit for additional learning. In effect, on the basis of their test scores, the former students would be classified as 'masters' of the learning materials in the current unit while the latter students would be classified as 'non-masters'. The ability of criterion-referenced tests to facilitate these kinds of decision making have made them attractive assessment devices in situations where it is desirable to make assessments tailored to the abilities and needs of individuals.

In addition to evaluating students' progress in objectives-based instructional programs, criterion-referenced tests are also being used to diagnose learning deficiencies, to evaluate the effectiveness of

various educational and social programs, and, what is now becoming an important use at the local, state and national levels, to assess competencies of individuals in basic 'societal' skills of numeracy and literacy.

During the past few years an increasing number of educational measurement specialists have shown a keen interest in the problem of correctly assigning examinees to one of several mutually exclusive mastery states, or categories, on the basis of the examinees' raw scores on a criterion-referenced test. For each raw score on the test it has been possible to estimate a measure of examinee performance for the domain of tasks or behavioural objectives being tapped by the test items. The thrust of this problem has been directed toward comparing each examinee's estimated performance with a specified standard of performance (often called a criterion score or minimum competency level), and not, as is usually the case with norm-referenced tests, of comparing each examinee's performance with the mean or other similar measure of the performance of the associated norm group (Livingston, 1972). In its simplest form the problem is focused on only two states - a mastery state characterizing those estimated performances for which examinee performance falls above the criterion level, and a non-mastery state characterizing those estimated performances for which examinee performance falls below the criterion level.

In order to interpret criterion-referenced test scores, the examiner must employ a measurement model which relates these scores to levels of performance for the behaviours being tapped by the test items. Since the introduction of criterion-referenced measurement into the educational measurement literature by Glaser and Klaus (1962), a variety of measurement models have been proposed by research workers to facilitate the interpretation of scores obtained with criterion-referenced tests.

Meskauskas (1976) has reviewed some of the more popular measurement models that have been proposed, at one time or another, for dealing with criterion-referenced test data, and for solving the mastery/non-mastery classification problem. His survey has shown quite conclusively that the nature of a criterion-referenced measurement model, and hence its usefulness in a practical setting, is related to how well the assumptions underlying the particular model agree with the interpretation of

the role of criterion-referenced measurement in estimating performance standards in that situation. A more recent work by Hambleton, Swaminathan, Algina and Coulson (1978) has provided an up-to-date review of the newer measurement models, reviewing at the same time important technical issues and developments in criterion-referenced testing and measurement.

Many articles have appeared in the literature of educational measurement examining various theoretical and practical aspects of criterion-referenced testing. Sumner and Robertson (1977) have summarized the findings of a large number of these studies, and have reviewed a number of criterion-referenced tests which have been or are currently being used as evaluation instruments.

Concepts of Criterion-Referenced Measurement and Tests

It appears that the first to use the term 'criterion-referenced testing' were Glaser and Klaus (1962), in their article on the assessment of human performance. In the following year, Glaser (1963), in the context of a discussion on instructional technology and the measurement of learning outcomes, elaborated in greater detail his conceptualization of criterion-referenced testing, defining it as a process of measuring an individual's achievement with respect to a body of knowledge and locating this measure on a continuum of knowledge ranging from a state of no proficiency at one end to a state of perfect performance at the other end. According to Glaser, each individual's score on a criterion-referenced test may be translated into a point on a continuum of knowledge acquisition, to provide explicit information as to the degree of competence in the underlying body of knowledge. A few years later, Popham and Husek (1969) published a paper that attempted to lay a distinctive conceptual framework for criterion-referenced measurement in contrast to that available for norm-referenced measurement. In contrasting norm-referenced and criterion-referenced measurement they said (Popham and Husek, 1969:2):

... norm-referenced measures are those which are used to ascertain an individual's performance in relationship to the performance of other individuals on the same measuring device Criterion-referenced measures are those which are used to ascertain an individual's status with respect to some criterion, i.e., performance standard.

Following these important pioneering works, numerous definitions and interpretations of criterion-referenced measurement and tests appeared from time to time in the educational measurement literature. Most definitions took the view that, at its heart, a criterion-referenced test is one that attempts to measure the status of an examinee with respect to some well-defined set or domain of tasks or behavioural objectives. Alkin (1974:3) has reviewed some popular definitions, relating the various terms that have been used to define criterion-referenced measurement. He found these definitions to differ considerably, especially in the 'limitations and constraints' placed on such measurement. However, most of these have emphasized the need to adopt systematic procedures when selecting items to be included in tests which are to provide measures of performance, and many definitions have also required that performance standards be set prior to the actual testing of examinees.

A definition of a criterion-referenced test that has found wide acceptance, is the one provided by Glaser and Nitko (1971:653), who said:

A criterion-referenced test is one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards. Performance standards are generally specified by defining a class or domain of tasks that should be performed by the individual. Measurements are taken on representative samples of tasks drawn from this domain, and such measurements are referenced directly to this domain for each individual measured.

In this respect, Block (1971) has asserted that criterion-referenced measurement involves the production of absolute indices of performance that are designed to indicate what an examinee has mastered (learned).

The representative sample of tasks in the Glaser-Nitko definition refer to a set of test items sampled from a well-defined population of items that relate in some definite way to the tasks in the domain. Kish (1965:82) has defined 'representative sampling' as:

... samples which are 'minatures of the population', and by the notion that the 'defined parts of the population should be appropriately represented in the sample'.

That is, if a certain domain were to consist of a number of distinct behaviours, and a test were to be constructed by representative sampling, then the test items would have to be sampled from the

population of items so as to adequately represent these behaviours in the test. If the sampling were not representative, it would not be possible to make precise statements about the actual meaning of the measured performances in terms of the behaviours in the domain. Indeed if the sampling were inadequate some behaviours might not be represented at all. Moreover, even now there are some measurement specialists who are skeptical that a precise, unambiguous relationship can exist in an operational sense, between the domain of tasks and the corresponding population of items (Alkin, 1974:6, Popham 1974:14). If this skepticism is well founded there could well be important implications for test developers; it may not be possible to determine with absolute certainty that a level of performance measured by the test items represents a corresponding level of performance on the domain of tasks. Furthermore there is the problem of constructing item populations to represent specific objectives. In this regard Harris (1974a:86) has cautioned:

For the test developer an important question is whether or not it is reasonable to postulate a population of similar items for any specific objective. The critical word is 'reasonable'. If the answer is no, then the developer is forced to use a single-item test to measure that objective. (He may, of course, decide to aggregate several such single-item tests rather than score them separately, but this does not alter the fact that he cannot sample from a defined population of items to measure that specific objective.)

The degree of item similarity mentioned by Harris is essentially a concept of item homogeneity/heterogeneity. Davis and Diamond (1974:118) have defined a set of items to be homogeneous if they all measure the same variable (behavioural objective), plus error, whereas they have defined a set of items to be heterogeneous if they all measure entirely uncorrelated variables (behavioural objectives), plus random error. It is likely that the more homogeneous a set of items, the less widespread are the item difficulties; so that items measuring a specific behaviour would have similar difficulties, provided, of course, the error of measurement was small for each item. Although the extreme cases of complete homogeneity and complete heterogeneity are not usually found in practice, they are not precluded by the Glaser-Nitko definition. This definition allows the possibility, for example, that a population of items consists of sets of heterogeneous items which are related to a domain of diverse and seemingly unrelated tasks. This view is

consonant with the view adopted by Millman (1973, 1974) who suggested that behavioural objectives (tasks) may be formed into a single narrow dimension or they may be formed into multidimensional ones, but in either case, before attempting to measure performance standards, the structure of the domain should always be precisely defined. In any case, Davis and Diamond (1974:119) have warned that:

As the degree of heterogeneity increases, total scores derived from such a test become more and more inappropriate for diagnostic purposes. But they may be entirely appropriate for survey purposes to estimate the proportion of a population of items constituting a domain, or area of subject-matter content, that can be answered correctly by an examinee.

Consequently, according to the Glaser-Nitko definition, a criterion-referenced test may be constructed by drawing a representative sample of items from a population of items that relate to either a limited number of tasks or to a wide spectrum of tasks. Therefore a test may be used to assess the global performance of an individual over all the dimensions of behaviours constituting the domain. If more refined measures of performance are needed the status of the individual should be determined separately for each dimension. In that case the multidimensional item population would first have to be partitioned into the requisite number of sub-populations, determined by the dimensionality of the test, before drawing representative tests from these sub-populations. Of course the degree of partitioning and thus the degree of item homogeneity within sub-populations would depend on the judgment of the test developer during the test construction phase.

Some writers have taken a somewhat different view of criterion-referenced measurement, particularly where heterogeneity (and item difficulty) is concerned. For example, Emrick (1971), Besel (1972), Macready and Merwin (1973), Besel (1973), Roudabush (1974), and Macready and Dayton (1977) have developed measurement models which postulate item-populations that are essentially homogeneous. In these models, which have come to be known as 'state models', examinees are assumed to respond consistently to all items sampled from the well-defined population of items. That is, these models assume that examinees are either masters of all items in the population or not at all. But, if an examinee who is a master does not answer correctly all the items in a test, it is supposed that the examinee's performance

is affected by extraneous factors which are represented as errors of measurement. The salient features of state and continuous models have been examined and contrasted by Meskauskas (1976).

Measurement of Mastery

The Glaser-Nitko conceptualization of measurement using criterion-referenced tests lends itself to the formulation of a class of evaluation models that view the ability of the examinee as lying along a continuum (Meskauskas, 1976). In such models the examinee's true ability is represented by a variable which measures the proportion of items in the population of items related to the domain of tasks, that an examinee could answer correctly if he were administered all the items in the population. This variable is restricted to take a set of values in a defined range, usually from 0 to 1.

Obviously the true ability of an examinee is an ideal measure that exists as a theoretical construct. It may be estimated to any degree of accuracy by taking a sufficiently large sample of items from the population. But the true ability is in fact a surrogate for the more fundamental measure of examinee performance for the domain of tasks. As Davis and Diamond (1974:122) have said:

... it is clear that no wholly adequate practical procedure is available, or can be made available, for estimating the proportion of tasks in a parent population of tasks that can be performed by an examinee; we can, strictly speaking, infer only what proportion of items in a parent population of items he can answer correctly unless the tasks themselves can be used as items.

In order to make mastery/non-mastery decisions, a mastery (criterion) level is frequently defined, prior to testing, as a value of this variable customarily near the top of the interval scale. Masters are those examinees whose true ability is equal to or exceeds the defined mastery value, whereas non-masters are those examinees whose true ability is less than the stated value (Novick and Lewis, 1974:147).

At present there is no one universally satisfactory method for arriving at a standard of performance (minimum competency) for a domain of behaviours or the corresponding population of items (Meskauskas, 1976; Glass, 1978). Indeed all methods must ultimately rely on the subjective judgment of the person or persons who will have

to interpret criterion-referenced test scores to determine what is and what is not a satisfactory level of performance. Some people, such as Glass (1978) and Burton (1978) have attacked the usefulness of criterion-referenced testing for making mastery/non-mastery decisions because the procedure involves, as they see it, the setting of standards in an arbitrary way; by simply moving the standard along the achievement continuum one could capriciously determine which scores will yield masters and which will yield non-masters. Even worse, these persons feel, is the likelihood that in certain situations independent judges may come up with different standards for the same set of behaviours. Proponents of criterion-referenced testing, such as Hambleton (1978), Block (1978) and Popham (1978), have responded to these criticisms by pointing out that all forms of testing involve judgment at some stage of their application, and that where judgments are made they are the result of careful and considered thought and not the result of random or whimsical actions as Glass and others would believe. Moreover, Popham (1978) has pointed out, that where a group of independent judges come up with different standards it may be that they had conceptualized the domain of tasks differently and/or adopted different standard setting procedures. Thus he said (Popham, 1978:298):

Different [standard setting] procedures can yield different results. What must be judged is which one of the different procedures is more appropriate.

So great and volatile is the debate on the setting of standards that recently a whole issue of the Journal of Educational Measurement (Winter 1978) has been devoted to this controversial subject, presenting arguments from a number of eminent educational measurement specialists who are for or against the use of standards of performance and the setting of minimum competency levels.

Bayesian Formulation of Criterion-Referenced Measurement

A number of educationalists (e.g. Cronbach and Gleser, 1965; Hambleton and Novick, 1973) have viewed all types of testing as essentially a decision process. However, the outcomes of decision processes are not the same for all types of testing. In criterion-referenced testing the decision process primarily aims to determine whether an individual has exceeded the cutting (or passing) score on the test, whereas in norm-referenced testing the decision process

primarily aims to select a certain proportion of individuals who have been ranked on the test. Because decision making is seen to be the essence of testing, some researchers are convinced that it is desirable to use only those inferential procedures for making inferences about individuals' abilities, which consider or are capable of considering consequences of taking particular actions at various stages of the testing process. Such inferential procedures, they believe, allow specification of a chain of optimum actions, that can be tested against data and alternative hypotheses if necessary (Novick, 1973).

Perhaps the ultimate goals of the examiner when testing students with criterion-referenced tests, are to give point or interval estimates of students' abilities, and to assign students to mastery or non-mastery states. Inferential procedures that have been found to be particularly useful for these purposes, are those based on the decision-theoretic methodology. Unfortunately, Bayesian statistical methods are not widely known in the educational community, and to date these evaluation models have had little impact on educational practice. Nevertheless some see that this methodology does have much to offer, especially when it is applied in a field as complex as mental testing, where the available test data is often too sparse for making acceptable inferences within the classical statistical framework.

Many of the important issues surrounding these inferential procedures, as well as the criterion-referenced measurement models, have been reported in the papers of Hambleton and Novick (1973), Novick, Lewis and Jackson (1973), Davis, Hickman and Novick (1973), Novick and Lewis (1974), Swaminathan, Hambleton and Algina (1975), Huynh (1976), and Hambleton, Swaminathan, Algina and Coulson (1978). It is worth noting that the Bayesian approach to criterion-referenced testing is currently finding most of its application in objectives-based instructional programs, educational programs, as was pointed out earlier, that are maximally adaptive to the requirements of individual learners (Hambleton, 1974; Spinetti and Hambleton, 1977).

The Bayesian decision-theoretic methodology combines Bayes' theorem with facets of statistical decision theory. In short, the Bayesian part of the procedure supplies a point or interval estimate of the examinee's ability, which may be used as an element in a decision problem. Indeed, a reason for applying this methodology to analyse criterion-referenced test data, is its ability to incorporate prior knowledge of examinee

performance while at the same time take into consideration opportunity losses arising from the possible misclassification of examinees. A succinct description of this methodology has been given by Box and Tiao (1973:19) who said:

... it is supposed that a choice has to be made from a set of available actions ($a_1, \dots, a_r$), where the payoff or utility [negative loss] of a given action depends on a state of nature, say θ , which is unknown. The decision maker's knowledge of θ is represented by a posterior distribution which combines prior knowledge of θ with the information provided by an experiment, and he is then supposed to choose that action which maximizes the expected payoff [minimizes the expected opportunity loss] over the posterior distribution.

There are a number of excellent books which provide a thorough coverage of various facets of this methodology. Thus Lindley (1965a; 1965b) and Box and Tiao (1973) have examined most aspects of Bayesian inference, giving sound mathematical treatments. Raiffa and Schlaifer (1961), Pratt, Raiffa and Schlaifer (1965), and De Groot (1970) have provided equally rigorous treatments of decision theory, particularly that associated with extensive and normal form analyses, which have been found useful for setting-up decision problems concerned with the determination of cutting scores for criterion-referenced tests.

For many measurement specialists, a point of contention with Bayesian statistics is that different decisions are likely to arise if different prior distributions are specified for examinee ability. According to Box and Tiao (1973), what is important, however, is that if priors disagree significantly in their specification of ability, and consequently lead to different actions, the option remains for recourse to more experimentation to resolve this impasse. Indeed they have claimed that in many applications of Bayesian statistics, slight differences in priors had not led to different actions within the one problem.

Sometimes, when applying Bayesian statistics to analyse criterion-referenced test data, the examiner may not possess useful prior information on examinee ability and, in this situation he may be forced to use a non-informative or uniform prior distribution to describe the lack of this information. Alternatively the examiner may use the collateral (group) information provided by the scores of all examinees on the test, in lieu of any specific knowledge of the examinee beyond

his test score. That is, the examinee's prior distribution could be identified as being equivalent, in a statistical sense, to the distribution of test scores of the group to which he belongs (Novick, Lewis and Jackson, 1973; Huynh, 1976).

Extraneous Variables in Criterion-Referenced Measurement

It is well documented in the literature of educational measurement that a particular set of response data from an examinee taking an objective test is the result of many overt and covert processes in addition to the ability being measured. Test response data are a function of the interaction of examinee and test. For the test to be considered a valid instrument for measuring the ability of interest, this interaction should produce data which essentially reflects the ability being measured.

But numerous studies have concluded that test response data are unlikely to be directly related to the ability being measured; these data are more often than not contaminated with the effects of many extraneous variables which are often represented by error terms in subsequent analyses. By evaluating results reported in the available literature, Cronbach (1946; 1950) showed that a significant portion of these extraneous variables can be identified as response sets of examinees. These response sets are personality characteristics of the examinee that produce consistency in the manner of response, within or across testings, to aspects of test form apart from the specific item or test content.

Cronbach had suggested that response sets operate so that individuals with the same level of true ability do not necessarily receive the same test scores. That is, response sets confound the effect of ability on observed test score. In fact, Cronbach had indicated that response sets may either increase or decrease the empirical validity and reliability of test scores. Response sets, apart from the test instrument, were therefore seen to contribute a significant portion of error of measurement and, where possible, should always be avoided or controlled.

Of the many response sets considered by Cronbach, two may be particularly important sources of measurement error in objective testing. They are the tendency to gamble or guess and the tendency to speed or

be careless. Much has been written on guessing and its ramifications for test score interpretation, particularly in norm-referenced testing; Diamond and Evans (1973) have provided a useful review of this subject.

Davis and Diamond (1974:121) have discussed, in terms of criterion-referenced testing, the influence of guessing on the determination of the true proportion of items in the population of items which the examinee knows. They applied the typical correction for guessing formula after cutting scores were determined using the Bayesian methodology. Recently Shaycoft (1976) had corrected continuum estimates of ability on a criterion-referenced test by incorporating the correction for guessing within the measurement model itself, rather than applying these corrections in a post-hoc fashion. Although she had successfully corrected for random guessing, her method is cumbersome since it involves calculating a number of probabilities that are arranged into matrices to be multiplied together to obtain estimates of the true scores.¹ Little else appears to have been published on criterion-referenced continuum models that correct for the effects of guessing.

The literature on carelessness in objective testing, however, is relatively sparse compared with that on guessing. A number of writers (Davis and Diamond, 1974:136; Emrick, 1971; and Macready and Dayton, 1977) have considered the effect of carelessness, mostly in terms of state models of criterion-referenced measurement. Emrick referred to this response set as 'careless error', and he saw this type of error as item responses which do not correspond to the examinee's true mastery status.

Recently Macready and Dayton (1977), using non-Bayesian methodology, have created a probabilistic state model which relates probability of occurrence of each possible response pattern to the proportion of individuals who are masters or non-masters, taking into account probabilities of 'guessing' and 'forgetting' (carelessness). The 'guessing' and 'forgetting' parameters in that model were estimated from the response pattern of the total group of examinees taking the test, using the maximum likelihood estimation procedure.

¹ The results presented by Shaycoft are more readily derived using a special case of the model presented in the current work.

Origin and Purpose of the Present Study

The work reported here had its genesis in the Australian Studies in School Performance, a project which was recently completed by the Australian Council for Educational Research (Bourke and Keeves, 1977). A prime objective of that study was to determine the proportion of 10- and 14-year-old Australian students who had mastered basic skills considered to be crucial for their effective functioning as literate and numerate persons in Australian society.

To obtain measures of each student's acquisition or non-acquisition of these basic skills new criterion-referenced tests were constructed according to a definite plan of objectives. Criterion-referenced tests were deemed to be more suitable for this purpose than traditional norm-referenced tests. A compelling reason for not using norm-referenced tests lay in their method of construction- these tests are generally constructed so as to maximize variability of students' scores over the score range that is of interest. An item is selected for inclusion in the test if it adds to the discriminating power of the test, and often little attention is paid to the specific nature of the behaviours underlying this item. On the other hand, in constructing criterion-referenced tests, items are selected from a well-defined population of items that relate to a set of behaviours, and little attention is paid to the items' discriminating powers.

The adoption of criterion-referenced tests for making mastery decisions concerning the students acquisition or non-acquisition of the basic skills required, in the first instance, the determination of cutting (passing) scores. The cutting scores would yield results, it was hoped, that would allow comparison of the proportions of students within and between the two age levels who had mastered the skills being assessed.

Simply reporting simple proportions of the various raw score distributions was viewed to be undesirable because they would give inaccurate estimates, primarily because many of the distributions were highly skewed and many of the sub-tests were short, some consisting of only four items. Furthermore, the accuracy of such simple estimates would probably have been affected considerably by biasing errors, such as those arising from the response sets of examinees.

These perceived difficulties suggested that a criterion-referenced measurement model should be developed which would at the very least provide a means for correcting for the effects (biasing errors) of two major response sets, namely guessing and carelessness. Such a model, it was anticipated, might be usefully incorporated into a Bayesian beta-binomial, decision-theoretic framework. A Bayesian approach, it was felt, would allow more accurate mastery/non-mastery decisions to be made, given the shortness of many of the tests, than could be possible if only simple proportions were used. Several versions of the measurement model were examined during the course of this study, and the one which was finally adopted assumed that true ability of an examinee is an unobserved quantity (considered to be a latent variable), but that it is related to observed ability (also a latent variable), which is true ability confounded by guessing and carelessness. This relationship was expressed by a linear equation in which guessing and carelessness were defined to be fixed parameters. This measurement model was not only able to provide cutting scores for each test, but allowed cutting scores to be determined for different levels of biasing error, represented by these two parameters.

In short, the aims of the current study, then, were to:

- 1 develop a fixed parameters model of criterion-referenced measurement, in which the guessing and carelessness parameters would be fixed to certain values when performing calculations;
- 2 incorporate this model into a beta-binomial framework of the Bayesian decision-theoretic methodology;
- 3 develop a computer program based on this methodology, which would facilitate the analysis of criterion-referenced test data;
- 4 examine the performance of this model in a computer simulation, observing the effects of different specifications of the parameters and other elements such as test length, prior distribution, criterion level and ratio of the opportunity losses; and finally
- 5 report examples of calculations that were used to determine cutting scores for tests used in the Australian Studies in School Performance.

CHAPTER 2

A MODEL OF CRITERION-REFERENCED MEASUREMENT WITH CORRECTIONS FOR GUESSING AND CARELESSNESS

In specifying a criterion-referenced measurement model it would seem essential to first define the structure of the domain of tasks on which measurements are to be made, and to define the structure of the corresponding population of items related to these tasks. The logical validity of a criterion-referenced measurement model depends upon the assumptions underlying these two constructs. Changed conceptions of either the structure of the domain of tasks or the structure of the population of items might very well lead to different conclusions about mastery, and hence the classification of examinees into mastery states. The present study develops a measurement model based on the definition of criterion-referenced measurement given by Glaser and Nitko (1971).

It is assumed an examinee's true ability is represented at a certain point in time, and with respect to a well-defined domain of tasks, by the continuous variable θ_t , which provides a representative measure of the proportion of items of the hypothetical population of items that the examinee has mastered. This view of ability allows for the possibility that there are some 'partly tutored students' (Harris, 1974b:102). That is, θ_t characterizes examinees with different backgrounds of experience concerning the tasks in the domain, so that students' abilities are likely to vary in proportion from 0 to 1, rather than be bunched (as in a state model of ability) at the bottom (0) or top (1). Students for whom $\theta = 1$, are called masters of the whole population of items, while students for whom $\theta = 0$, are called non-masters of this population of items. Students for whom $0 < \theta_t < 1$, are called masters of some items and non-masters of other items. According to Harris, θ_t is the conditional probability of passing a randomly chosen item from the population given that another item has been passed. Of course this approach assumes the interchangeability of items for the purpose of defining θ_t . In the continuum model of criterion-referenced measurement it is assumed that the conditional probability may be any value from 0 to 1.

Binary Response Test Data

Consider the hypothetical population of items such that an item sampled from this population is scored 1 if an examinee answers it correctly, but is scored 0 otherwise. According to Lord and Novick (1968:362), every examinee taking a test consisting of n randomly chosen items from this population may be characterized by an n -dimensional response vector $\underline{v} = (u_1, \dots, u_g, \dots, u_n)$, where u_g is the response to the g th item of the test.

The ability of an examinee with respect to each item can be stated in terms of a conditional probability. Thus the conditional probability of observing a particular item score ($u_g = 1$ or $u_g = 0$) given the true ability θ_t of the examinee, and given a number of item parameters representing factors other than ability, is represented by the variable θ_{og} called the observed ability of the examinee on item g . Mathematically, this may be written

$$\begin{aligned} (2.1) \quad \theta_{og} &= \Pr(u_g = 1 \mid \theta_t, \underline{a}_g) \\ &= 1 - \Pr(u_g = 0 \mid \theta_t, \underline{a}_g) \end{aligned}$$

where $\underline{a}_g$ is a vector of item parameters.

Probabilistic equations such as (2.1) have often been plotted as item characteristic curves, and they form the starting point in the development of most latent trait models of mental tests (Lord and Novick, 1968:358; Lord, 1977; Hambleton and Cook, 1977).

The item response score u_g may be considered to be the realization of a random variable U_g . It is assumed this variable is associated with a distribution of item scores arising from repeated administrations of the g th item to the one examinee. However it is possible to make the assumption, although this is not made here, that the random variable U_g describes the distribution of item scores across a particular group of examinees with respect to the g th item. This kind of assumption would be necessary if inferences were to be drawn from the distribution of ability across this group of examinees.

Hence the distribution of item scores for one examinee was taken to be

$$(2.2) \quad f_g(u_g \mid \theta_t, \underline{a}_g) = \Pr(U_g = u_g \mid \theta_t, \underline{a}_g) .$$

This equation may be written

$$f_g(u_g \mid \theta_t, \underline{a}_g) = \theta_{og}^{u_g} (1 - \theta_{og})^{1-u_g}$$

where $u_g = 0$ or 1 .

If there are n items in the test, the probability of obtaining a particular pattern of responses described by the response vector

$\underline{v} = (u_1, \dots, u_n)$, is

$$\Pr(\underline{V} = \underline{v} \mid \theta_t, \underline{a}) = \Pr(U_1 = u_1, \dots, U_n = u_n \mid \theta_t, \underline{a}),$$

where $\underline{a} = (\underline{a}_1, \dots, \underline{a}_g, \dots, \underline{a}_n)$ is the vector of item parameters for the total test. If local independence of items is assumed (Lord and Novick, 1968:360) the above identity becomes

$$(2.3) \quad \Pr(\underline{V} = \underline{v} \mid \theta_t, \underline{a}) = \Pr(U_1 = u_1 \mid \theta_t, \underline{a}_1) \dots \Pr(U_n = u_n \mid \theta_t, \underline{a}_n)$$

$$= \prod_{g=1}^n \theta_{og}^{u_g} (1 - \theta_{og})^{1-u_g}$$

The Model: Relation Between Observed Ability and True Ability

Equation (2.1) indicates the observed ability of an examinee, measured by the g th item in a test, is a probability which is governed by the examinee's true ability and item parameters representing effects distinct from true ability. In the present study, a particular expression of this probability was derived for testing situations considered to involve only two kinds of systematic error, namely that resulting from guessing and carelessness. In effect, the derived functional relationship between observed ability θ_{og} and true ability θ_t was specified in terms of a guessing parameter β_g , which is the probability that the examinee selects the correct answer given that he guesses, and a carelessness parameter γ_g , which is the probability that the examinee is careless when responding to the item.

To ascertain the probability with which the examinee guesses the correct answer, given that he guesses, is a difficult if not impossible task. However, the task may be simplified by assuming β_g is determined by the particular item structure of the test. That is, should the test consist of multiple-choice items, with M_g alternatives per item, this assumption would have $\beta_g = \frac{1}{M_g}$. Of course this implies the examinee guesses randomly among the alternatives, which are supposedly equally attractive to him. In general, however, examinees do not guess randomly but guess on the basis of their partial knowledge, so that in a multiple-choice item some alternatives could be eliminated by the examinee as not being plausible, and consequently the examinee would guess from among the remaining alternatives. Nevertheless, in developing a model which relates true and observed ability, random guessing was taken to be a meaningful assumption for the purposes of the present study, even though it may be a bold one in some applications outside this study. It is worthwhile noting that the assumption of random guessing is quite common in the field of mental testing mainly because researchers have found the general case far too complex (Lord and Novick, 1968: 302-318).

As examples of random guessing, consider multiple-choice items with $M_g = 4$ alternatives per item, and completion type items for which the examinee has to provide his own answer. In the former example $\beta_g = 0.25$ and in the latter $\beta_g = 0$. Completion type items may be considered as an extreme case of the multiple-choice item, in which the number of alternatives per item is, theoretically speaking, infinite. The probability of guessing the correct answer, given that the examinee guesses, is for this type of item very small indeed.

The probability that the examinee will be careless when responding to the g th item is equal to γ_g . A reference to a dictionary will reveal that 'carelessness' is a broad concept that has different meanings in different contexts. Here it is taken in its broadest sense to mean a process that describes the examinee's response as being unconcerned or light-hearted or thoughtless or negligent and, beyond this, no attempt is made to justify a more precise meaning or to give a fully operational definition. This broad definition allows the

examinee to be careless independently of the structure of the test, in so far as it may be taken to be a useful global indicator of the examinee's tendency to make errors when responding to an item. It is assumed to function independently of the examinee's true ability. High ability students are just as liable to be careless as low ability students and conversely; in practice, γ_g will not only vary across a population of examinees but it will also vary across items for a specific examinee.

Although carelessness is defined as a process of responding in which the examinee is likely to make errors, this definition does not preclude the possibility that his response may culminate with the selection of the correct answer. Indeed, it is well known by test developers that in certain test programs which are not part of the usual classroom situation, many examinees respond to items in an unconcerned and haphazard manner or respond to items after expending some minimal amount of light-hearted effort. These students are often not fully motivated because they do not perceive any benefits accruing to them from the test program or they do not see any value in the test instrument itself. What is not surprising, though, is that even if such students are observed to be careless and they are known to possess the requisite knowledge to answer the items correctly, they can still manage to obtain correct answers apparently without having used their knowledge. Informal evidence such as this suggests that examinees with the requisite knowledge but who respond carelessly to items may still manage to obtain the correct answers, especially in multiple-choice tests.

For example, suppose the examinee was unconcerned and that he responded in a haphazard fashion. Suppose also that he possessed the requisite knowledge to derive the answer and that he selected the keyed alternative apparently without using this knowledge or without expending any meaningful effort. What criteria might have guided him to help him select the keyed alternative is not known. In fact all the examiner can presume, after the event, is that the examinee responded without using his knowledge but instead used some selection criteria known only to himself. Indeed his obtaining the correct answer, given that the examiner knows he is completely careless, suggests that in this situation the examinee responded as though he had guessed randomly. Had the examiner known beforehand that the examinee would

respond like this he could have put $\gamma_g = 1$.

In contrast, suppose the examinee's response was not entirely haphazard but the result of some light-hearted effort, which, however, was insufficient to enable the derivation of the correct answer. In this case the examinee's response was inefficient because he had not utilized all the requisite knowledge to solve the problem, which he is supposed to have. Suppose also that he was able to choose the correct answer. One plausible explanation for this apparently illogical result is that once the examinee had progressed a few steps through the problem; he stopped because the going became tough and then had made a choice from among the alternative answers using some selection criteria known only to himself. It would appear as though he had guessed, but this time possibly on the basis of his partial solution of the problem. Had this examinee been sufficiently motivated and concerned with his effort in the first place, the probability of him being careless would have been close to zero, and the examiner knowing this could have put $\gamma_g = 0$.

In deriving the model which relates observed ability to true ability, the following events were assumed to be exclusive and exhaustive, and to completely represent the examinee's interaction with the g th item in the test.

They are:

E = {examinee knows the correct answer and is not careless}

F = {examinee knows the correct answer and is careless}

G = {examinee does not know the correct answer}

By the multiplication law of probabilities (Lindley, 1965a:14), these events occur with probabilities:

$$\Pr(E) = \theta_t(1 - \gamma_g),$$

$$\Pr(F) = \theta_t\gamma_g,$$

$$\Pr(G) = 1 - \theta_t.$$

Furthermore it was assumed that

$$(i) \quad \Pr(u_g = 1 \mid E) = 1$$

and

$$(ii) \quad \Pr(u_g = 1 \mid \text{not-}E) = \beta_g = \frac{1}{M_g},$$

where $\text{not-}E = F \cup G$.

The conditional probability (i) states that the examinee will obtain an item score equal to 1, provided he knows the correct answer and is not careless when responding to the item.

The meaning of the conditional probability (ii) can be considered in terms of the separate events F and G, because $\text{not-}E = F \cup G$, and because F and G are mutually exclusive. If event G occurs (ii) becomes $\Pr(u_g = 1 \mid G) = \beta_g$. This conditional probability means: given the examinee does not know the correct answer, then if he obtains an item score equal to 1, it will seem as though he had guessed randomly among the M_g alternatives. If the event F occurs (ii) becomes $\Pr(u_g = 1 \mid F) = \beta_g$. This conditional probability means: given the examinee does know the correct answer and is careless when responding to the item, then if he obtains an item score equal to 1, it will seem as though he had guessed randomly among the M_g alternatives. Hence when either event F or event G occurs it is assumed that the examinee responds as though he were guessing randomly among the M_g alternatives. In the light of the above discussion on β_g and γ_g , this assumption will at best yield an approximate model that corrects for guessing and carelessness.

Now (2.1) is

$$\theta_{og} = \Pr(u_g = 1 \mid \theta_t, \underline{\alpha}_g),$$

where $\underline{\alpha}_g = (\beta_g, \gamma_g)'$. But this can be written as

$$\begin{aligned} \theta_{og} &= \Pr(u_g = 1 \text{ and } E) + \Pr(u_g = 1 \text{ and not-}E) \\ &= \Pr(u_g = 1 \mid E)\Pr(E) + \Pr(u_g = 1 \mid \text{not-}E)\Pr(\text{not-}E) \\ &= \Pr(u_g = 1 \mid E)\Pr(E) + \Pr(u_g = 1 \mid \text{not-}E)(1 - \Pr(E)) \end{aligned}$$

Substituting the probabilities for the events E, F and G, gives

$$\theta_{og} = 1.\theta_t(1 - \gamma_g) + \beta_g(1 - \theta_t(1 - \gamma_g)) .$$

This equation can be rearranged to give the simpler form

$$(2.4) \quad \theta_{og} = \beta_g + \theta_t(1 - \beta_g)(1 - \gamma_g) .$$

Equation (2.4) expresses the relationship between the observed and true abilities in terms of the guessing and carelessness parameters. Furthermore, $\beta_g \leq \theta_{og} \leq \beta_g + (1 - \beta_g)(1 - \gamma_g)$ because $0 \leq \theta_t \leq 1$. Also $0 \leq \beta_g \leq 1$ and $0 \leq \gamma_g \leq 1$.

Ideally the item parameters β_g and γ_g are free parameters, taking particular values for each examinee. Consequently when fitting the model to response data these parameters should be estimated for each examinee and each test item. However, in the present study no attempt was made to make such estimates, but rather the parameters were assigned a priori and assumed equal for all examinees and for all items in the one test. They were assigned values best thought to represent the testing situations. Hence the assumptions that $\beta_g = \beta = \frac{1}{M}$ and $\gamma_g = \gamma$ for all g, so that equation (2.4) became

$$(2.5) \quad \theta_o = \Pr(u_g = 1 \mid \theta_t, \underline{\alpha}) = \theta_{og} = \beta + \theta_t(1 - \beta)(1 - \gamma) .$$

Equation (2.5) may be interpreted as follows: In a population of 100 items, examinees with ability θ_t would be expected to obtain $100\theta_t$ items correct if they were not careless. If the probability that they are careless is γ , then the number of items they expect to get right is $100\theta_t(1 - \gamma)$. Of the $(100 - 100\theta_t(1 - \gamma))$ items remaining, they may be expected to get $\beta(100 - 100\theta_t(1 - \gamma))$ right by a process which is effectively one of random guessing, so that

$$100\theta_0 = 100\theta_t(1 - \gamma) + \beta(100 - 100\theta_t(1 - \gamma))$$

$$= 100\beta + 100\theta_t(1 - \beta)(1 - \gamma)$$

or

$$\theta_0 = \beta + \theta_t(1 - \beta)(1 - \gamma) .$$

Figure 2.1 shows plots of equation (2.5) for a four options multiple-choice test, for which $\gamma = 0, 0.05, 0.5, 1.0$, and $\beta = 0.25$. Provided $\gamma \neq 1$, the observed ability θ_0 increases monotonically with true ability θ_t . The dashed line represents the situation when

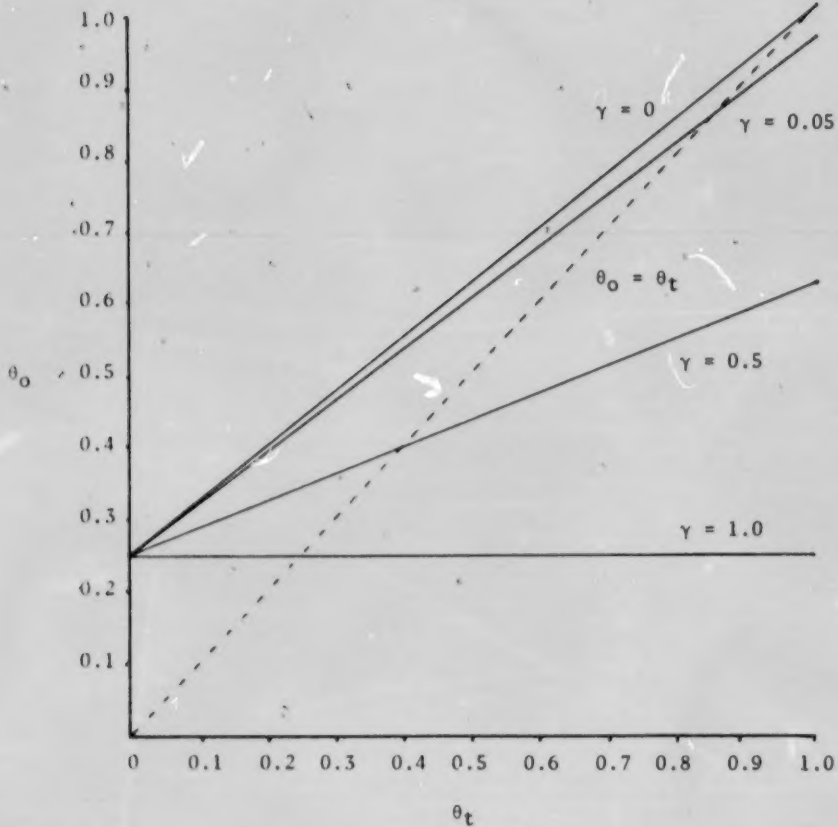


Figure 2.1 Plots of the Relation $\theta_0 = \beta + \theta_t(1 - \beta)(1 - \gamma)$ for Stated Values of γ and $\beta = 0.25$

the observed and true abilities are equal. Clearly this occurs when $\beta = \gamma = 0$; an ideal situation of no guessing and carelessness. In practice this situation might be approximated by a testing with constructed response items, where the examinee has to supply the correct answer. The line for which $\gamma = 1$ represents the case of complete carelessness, and is described by the equation $\theta_o = 0.25$. This equation seems plausible because, in a practical testing situation, complete carelessness may be viewed as random guessing on every item of the test. Thus in the four options multiple-choice test we would expect, under these circumstances, that the observed ability would take the value 0.25 irrespective of the true ability of the examinee. The three other lines in the figure represent situations of varying degrees of carelessness.

Above and to the left of the dashed line $\theta_o > \theta_t$, and below and to the right of the line $\theta_o < \theta_t$. In the former case the examinee's observed ability is inflated because there is guessing of some items in the test, and the inflation is more marked for an examinee whose true ability is small. In the latter case, however, the level of carelessness is more influential than guessing, resulting in a depression of the observed score. The true ability is equal to the observed ability for given values of β and γ when:

$$(2.6) \quad \theta_t = \frac{\beta}{(1 - (1 - \beta)(1 - \gamma))}$$

With $\beta = 0.25$ and $\gamma = 0.05$, equation (2.6) reduces to $\theta_t = 0.87$. Hence when $\theta_t > 0.87$, it follows that $\theta_t > \theta_o$; when $\theta_t < 0.87$, it follows that $\theta_t < \theta_o$. Obviously, β and γ interact, and the result of this interaction is related to the true ability of the examinee.

CHAPTER 3

A BAYESIAN DECISION-THEORETIC PROCEDURE FOR ESTIMATING MASTERY STATUS OF AN EXAMINEE

The Two-Action Decision Problem

Although it is sometimes useful to define more than two categories of mastery in criterion-referenced testing, the present study has chosen to restrict its attention to only two categories. Choosing two categories of mastery results in the two-action (classify the examinee as a master or classify the examinee as a non-master) problem. This problem may be considered to be solved for a particular individual taking a criterion-referenced test, if either action is selected in preference to the other, and the action that is selected is optimal in the sense that it minimizes expected opportunity losses (or disutilities) arising from possible misclassification of the examinee.

Now before an examinee can be assigned to a mastery state it is necessary to first define a criterion level on the population of items, against which the examinee's true ability may be judged. The criterion level defines what the examiner considers to be the minimum proportion of items in the population of items that the examinee should master before he can be accepted as a master of the related domain of tasks. Certainly the criterion level should be carefully chosen, and its value should be justifiable in terms of the objectives of the testing program. In the literature of criterion-referenced testing, values of this variable have usually been chosen to be in the range from 0.75 to 1.0 (Block, 1972).

In classifying the examinee, the examiner can make two kinds of classification errors depending on which one of two actions he takes. He may classify the examinee as a master when in fact the examinee is a non-master (meaning that his true ability is less than the criterion level), or he may classify the examinee as a non-master when in fact the examinee is a master (meaning that his true ability is greater than or equal to the criterion level). The former classification error is often called a false positive error, while the latter classification error is often called a false negative error (e.g. Huynh, 1976).

Table 3.1 Losses $L_t(d_i, \theta_t)$ Associated with Incorrect Actions
(θ_t metric)

		True ability	
		$\theta_t \geq \bar{\theta}_c$	$\theta_t < \bar{\theta}_c$
Decision	d_1 : Classify as a Master	0	$L_t(d_1, \theta_t)$
	d_2 : Classify as a Non-Master	$L_t(d_2, \theta_t)$	0

These situations can be conveniently represented in a table which reports opportunity losses that arise when either action is taken. Thus Table 3.1 reports in symbolic terms the kinds of losses that accrue when either a correct or incorrect decision is made.

In this table $\bar{\theta}_c$ is the true criterion level and $L_t(d_i, \theta_t)$ are functions that describe the magnitude of the opportunity losses for incorrect actions. For correct actions the loss function is assumed to be zero, as has often been assumed by those working in this field. In general the loss function depends on the examinee's true ability; and this dependence may be different for each action. Therefore if there is an explicit dependence on true ability, the magnitude of the loss can vary as the distance between the true ability and the criterion level is changed. Indeed the entries in any table such as Table 3.1 are based on subjective judgment of the examiner. In fact different examiners may postulate different loss functions for the same examinee because they may differ in their perceptions of the seriousness of the consequences arising from each action.

Table 3.2 Losses $L_o(d_i, \theta_o)$ Associated with Incorrect Actions
(θ_o metric)

		Observed ability	
		$\theta_o \geq \theta_c$	$\theta_o < \theta_c$
Decision	d_1 : Classify as a Master	0	$L_o(d_1, \theta_o)$
	d_2 : Classify as a Non-Master	$L_o(d_2, \theta_o)$	0

To cast the decision problem in terms of the criterion-referenced measurement model developed in the previous chapter, equation (2.5) can be used to transform the elements appearing in Table 3.1 to the θ_0 (observed ability) metric. Thus Table 3.2 represents in symbolic terms the equivalent opportunity losses in the θ_0 metric. In this table the observed criterion level θ_c , corresponds to the true criterion level $\bar{\theta}_c$; and it is obtained by replacing θ_t by $\bar{\theta}_c$ in equation (2.5). That is, the relation between θ_c and $\bar{\theta}_c$ is given by

$$(3.1) \quad \theta_c = \beta + \bar{\theta}_c(1 - \beta)(1 - \gamma)$$

Bayesian Solution of the Two-Action Decision Problem

As a reasonable approximation it is useful to assume that the loss functions shown in Tables 3.1 and 3.2, are constant for each region of the observed ability interval, thus implying a decision problem with threshold loss (Novick and Jackson, 1974:117). This means that the magnitude of the loss incurred is independent of the distance of the examinee's true ability from the criterion level. In other words, an examinee whose true ability is far from the criterion level and who is misclassified incurs the same loss as an examinee whose true ability is close to the criterion level and who is similarly misclassified.

Let $C_f = L_0(d_1, \theta_0)$ be the constant loss of granting mastery status to an examinee who is a failure (false positive error) and let $C_s = L_0(d_2, \theta_0)$ be the constant loss of granting non-mastery status to an examinee who is a success (false negative error) (Huynh, 1976). Both losses are here assumed to be non-negative real numbers that reflect the relative importance attached by the examiner to making each type of classification error.

Normally, within the foregoing decision-theoretic framework, the examiner is interested in taking the action that will yield the smaller expected loss, since he does not know the actual value of θ_t or θ_0 . Davis, Hickman and Novick (1973) (see also Novick and Lewis, 1974) show that this action is the one that gives the smaller value of the average posterior loss in an extensive form analysis. In the present work this has meant choosing the action (d_1 or d_2) which minimizes

in the θ_0 metric the Bayes risk function $r(x, d_1)$, defined by

$$(3.2) \quad r(x, d_1) = \int_{\beta}^{\beta + (1-\beta)(1-\gamma)} L_0(d_1, \theta_0) f(\theta_0 | x) d\theta_0.$$

In this equation $f(\theta_0 | x)$ is the posterior density function of θ_0 , given that x is the observed raw score of the examinee on the test. The posterior density function can be determined from an application of Bayes' theorem, which gives

$$(3.3) \quad f(\theta_0 | x) = \frac{g(x | \theta_0) \rho(\theta_0)}{\int_{\beta}^{\beta + (1-\beta)(1-\gamma)} g(x | \theta_0) \rho(\theta_0) d\theta_0},$$

where $g(x | \theta_0)$ is the likelihood function for the test, and $\rho(\theta_0)$ is a probability density function that indicates the examiner's belief as to the probable location of the examinee's observed ability in the interval $[\beta, \beta + (1-\beta)(1-\gamma)]$. In applications, $\rho(\theta_0)$ is often called the prior distribution of examinee ability or simply the prior.

The Bayes risk (or equivalently the expected loss) for action d_1 is

$$r(x, d_1) = \int_{\beta}^{\beta + (1-\beta)(1-\gamma)} L_0(d_1, \theta_0) f(\theta_0 | x) d\theta_0.$$

Because $\beta \leq \theta_c \leq \beta + (1-\beta)(1-\gamma)$ this risk can be written as

$$r(x, d_1) = \int_{\beta}^{\theta_c} L_0(d_1, \theta_0) f(\theta_0 | x) d\theta_0 + \int_{\theta_c}^{\beta + (1-\beta)(1-\gamma)} L_0(d_1, \theta_0) f(\theta_0 | x) d\theta_0.$$

But $L_0(d_1, \theta_0) = C_f$ when $\theta_0 < \theta_c$ and $L_0(d_1, \theta_0) = 0$ when $\theta_0 \geq \theta_c$.

$$\begin{aligned}\text{Consequently, } r(x, d_1) &= C_f \int_{\beta}^{\theta_c} f(\theta_0 | x) d\theta_0 \\ &= C_f \Pr(\theta_0 < \theta_c | x) .\end{aligned}$$

Similarly it can be shown that

$$\begin{aligned}r(x, d_2) &= C_s \int_{\theta_c}^{\beta + (1-\beta)(1-\gamma)} f(\theta_0 | x) d\theta_0 \\ &= C_s \Pr(\theta_0 \geq \theta_c | x) .\end{aligned}$$

Thus the expected loss for making a false positive error is given by $C_f \Pr(\theta_0 < \theta_c | x)$ and the expected loss for making a false negative error is given by $C_s \Pr(\theta_0 \geq \theta_c | x)$. Clearly then, the type of decision needed depends on the relative magnitudes of these two expected losses - the action associated with the smaller expected loss is the optimum action. If both expected losses have the same value then either action can be considered to be optimum.

In many instances, however, the examiner may want to know the score on the test, if it exists, where the two expected losses are equal. That is, he wants to find that score $x = c$, often called the cutting score, for which

$$C_f \Pr(\theta_0 < \theta_c | c) = C_s \Pr(\theta_0 \geq \theta_c | c) .$$

This equation can be written as

$$C_f(1 - \Pr(\theta_0 \geq \theta_c | c)) = C_s \Pr(\theta_0 \geq \theta_c | c)$$

or

$$(3.4) \quad \frac{C_f}{C_f + C_s} = \Pr(\theta_0 \geq \theta_c | c) .$$

The examinee can then be classified as a master if $x \geq c$, or as a non-master if $x < c$.

The Beta-Binomial Model

Before equation (3.4) can be used to calculate the cutting score c , it is necessary to calculate the value of the posterior probability $\Pr(\theta_0 \geq \theta_c \mid x)$. But this entails first finding the form of the posterior density function, $f(\theta_0 \mid x)$.

Suppose the test consists of a random representative sample of n items from some population of items. If the random variable X represents the total test score of an examinee on the test, then $X = x(\underline{v}) = x(u_1, \dots, u_n)'$ is a total test score formula, relating observed test score x to the response vector $\underline{v}' = (u_1, \dots, u_n)'$.

According to Birnbaum (1968:406) the cumulative distribution function of the test score can be written as

$$\begin{aligned} (3.5) \quad F(x \mid \theta_t, \underline{\alpha}) &= \Pr(X(\underline{V}) \leq x \mid \theta_t, \underline{\alpha}) \\ &= \sum_{X(\underline{V}) \leq x} \Pr(\underline{V} = \underline{v} \mid \theta_t, \underline{\alpha}) \end{aligned}$$

Now, if all items take the same values of the item parameters, and all items are equally weighted in the test, the test score formula is

$$x = \sum_{g=1}^n u_g.$$

Then from (2.3),

$$\begin{aligned} \Pr(\underline{V} = \underline{v} \mid \theta_t, \underline{\alpha}) &= \prod_{g=1}^n \theta_{og}^{u_g} (1 - \theta_{og})^{1-u_g} \\ &= (\theta_{og})^{\sum_{g=1}^n u_g} (1 - \theta_{og})^{n - \sum_{g=1}^n u_g} \\ &= (\theta_o)^x (1 - \theta_o)^{n-x}. \end{aligned}$$

Equation (3.5) is simply the binomial cumulative distribution function for n trials, with probability of success given by θ_0 . That is

$$(3.6) \quad F(x | \theta_0) = \sum_{i=0}^x \binom{n}{i} \theta_0^i (1 - \theta_0)^{n-i},$$

for $x = 0, 1, \dots, n$.

Thus the following is assumed to apply in criterion-referenced testing:

- 1 The result of the examinee answering each item in the test can be classified into one of two categories: success or failure.
- 2 The examinee can answer correctly a proportion θ_t of the population of items of which the test is a representative sample.
- 3 The response of an examinee to each item in the test, is independent of his responses to all other items in the test.
- 4 The test consists of a fixed number of items, say n .

Hence, from (3.6)

$$(3.7) \quad g(x | \theta_0) = \binom{n}{x} \theta_0^x (1 - \theta_0)^{n-x}, \quad x = 0, 1, \dots, n.$$

Equation (3.7) is the likelihood function which states the binomial probability density function for x correct answers given that the observed ability of the examinee is θ_0 .

Furthermore $\rho(\theta_0)$ is specified to take the form

$$(3.8) \quad \rho(\theta_0) = \frac{K}{B(a, b)} \theta_0^{a-1} (1 - \theta_0)^{b-1}, \quad \beta \leq \theta_0 \leq \beta + (1 - \beta)(1 - \gamma),$$

where K is a normalizing constant, and $B(a, b)$ is a complete beta function with parameters a and b . Equation (3.8) defines a class of prior density functions which have been assumed to take the form of a beta function with parameters a and b . A reason for choosing a beta function to represent the prior density is that the beta function combines simply with the binomial function. That is, it is naturally

conjugate to the binomial distribution. Moreover, because the beta function is a two parameter function, it has been found useful for generating a family of probability density functions that cover a wide range of prior beliefs in the distribution of the examinee's observed ability (Novick and Jackson, 1974:109).

Substituting equations (3.7) and (3.8) into equation (3.3) yields the posterior density function

$$(3.9) \quad f(\theta_o | x) = \frac{\frac{K}{B(a,b)} \binom{n}{x} \theta_o^{a+x-1} (1 - \theta_o)^{b+n-x-1}}{\frac{K}{B(a,b)} \binom{n}{x} \int_{\beta}^{\beta + (1-\beta)(1-\gamma)} \theta_o^{a+x-1} (1 - \theta_o)^{b+n-x-1} d\theta_o}$$

$$\text{or} \quad f(\theta_o | x) = \frac{L}{B(p,q)} \theta_o^{p-1} (1 - \theta_o)^{q-1},$$

$$\text{where} \quad \beta \leq \theta_o \leq \beta + (1 - \beta)(1 - \gamma),$$

$$B(p,q) = \frac{(p-1)!(q-1)!}{(p+q-1)!},$$

$$p = a + x, \quad q = b + n - x,$$

and

$$(3.10) \quad \frac{1}{L} = \frac{1}{B(p,q)} \int_{\beta}^{\beta + (1-\beta)(1-\gamma)} \theta_o^{p-1} (1 - \theta_o)^{q-1} d\theta_o.$$

Then the probability that θ_o is equal to or greater than the observed criterion level θ_c , posterior to obtaining the observed score x , is

$$\begin{aligned}
 (3.11) \quad \Pr(\theta_o \geq \theta_c \mid x) &= L \int_{\frac{\beta + \theta_c(1-\beta)(1-\gamma)}{\beta + \theta_c(1-\beta)(1-\gamma)}}^{\frac{\beta + (1-\beta)(1-\gamma)}{\beta + (1-\beta)(1-\gamma)}} \frac{1}{B(p,q)} \theta_o^{p-1} (1 - \theta_o)^{q-1} d\theta_o \\
 &= \Pr(\theta_t \geq \bar{\theta}_c \mid x) .
 \end{aligned}$$

The integrals in (3.10) and (3.11) may be evaluated either by numerical integration or by referring to tables of the incomplete beta function (Pearson, 1968). The cutting score, c , is defined to be that value of x which satisfies equation (3.4) when the probability expressed in (3.11) is used. However, since x is assumed to be an integer it is possible that no such x will exactly satisfy this equation. In this case an approximate solution may be obtained by choosing an integer value of x which yields the smallest difference between the terms on the left and right sides of the equality in equation (3.4).

CHAPTER 4

TRIAL CALCULATIONS: A COMPUTER SIMULATION OF THE MODEL

The trial calculations were intended to be exploratory. Their main purpose was to show by computer simulation the range of results that could be obtained from different specifications of test length, prior distribution, guessing and carelessness parameters, criterion level, and loss ratio.

Method

Calculations were restricted to test lengths of five and ten items. These test lengths were sufficient to indicate the effect of test length on the posterior probability and the cutting score. Program BAYES (Appendix C) was used to generate results based on elements listed below, which were entered into the calculations in various combinations.

Prior (Beta) Distribution of True Ability

(See Figure 4.1)

$n = 5$

$n = 10$

$B(1,1), B(3,2), B(4,1), B(4.5,0.5),$ $B(6,4), B(8,2), B(9,1)$	$B(1,1), B(3,2), B(4,1), B(4.5,0.5),$ $B(6,4), B(8,2), B(9,1), B(12,8),$ $B(16,4), B(18,2)$
---	---

Guessing and Carelessness Parameters

$(B, \gamma) = (0,0), (0,0.05), (0,0.20), (0.25,0), (0.25,0.05), (0.25,0.20)$

True Criterion Level

$\hat{\theta}_c = 0.70 \quad 0.75 \quad 0.80 \quad 0.85 \quad 0.90 \quad 0.95$

Loss Ratio

$C_s / C_f = 5 / 1 \quad 2 / 1 \quad 1 / 1 \quad 1 / 2 \quad 1 / 5$

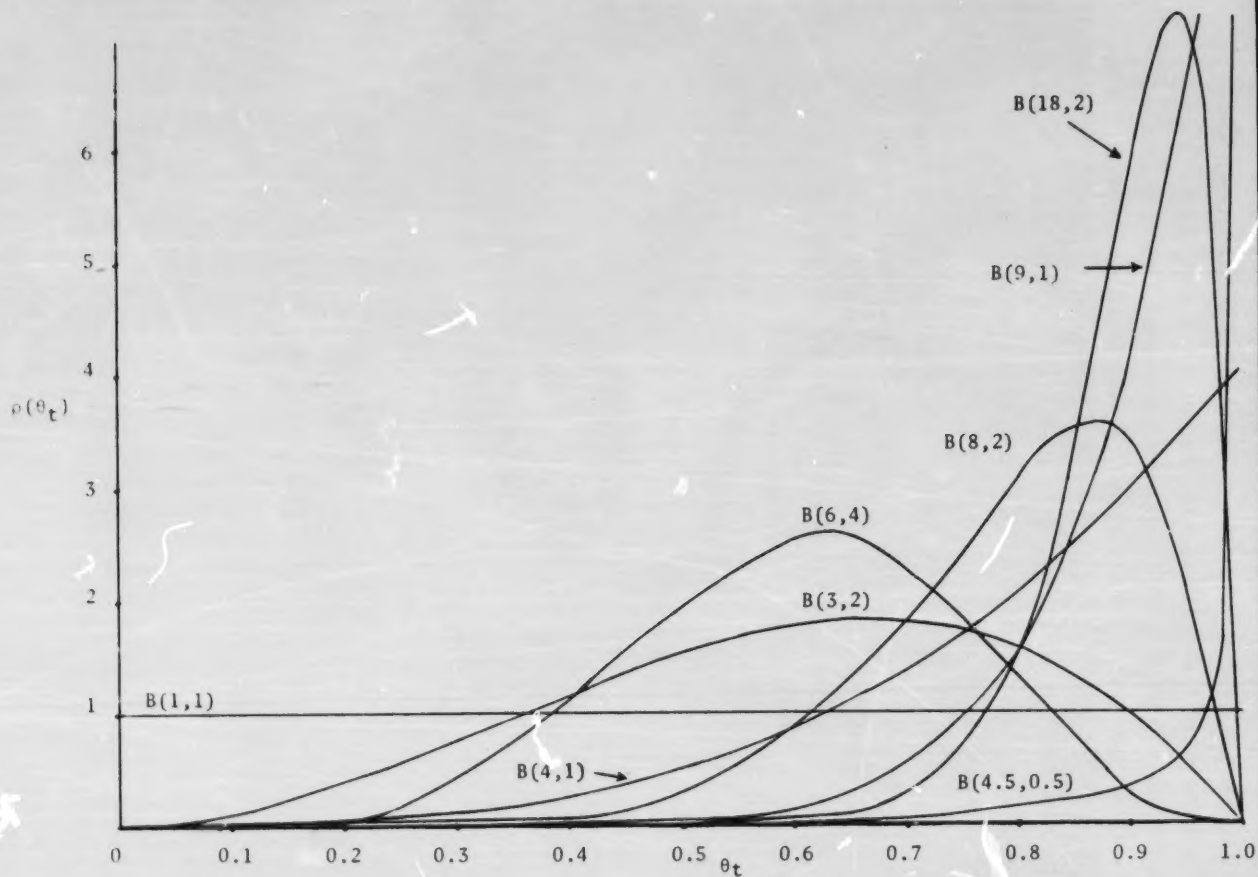


Figure 4.1 Examples of Graphs of $\rho(\theta_t) = B(a,b)$ where $B(a,b)$ is a Beta Function with Parameters a and b

Parameters of Prior Distributions

Since the true ability lies in the interval $[0,1]$ the corresponding observed ability must lie in the interval $[\beta, \beta + (1 - \beta)(1 - \gamma)]$. But $0 < \beta < 1$ and $0 < \gamma < 1$ and hence the observed ability interval is always the true ability interval or some constriction of it. Table 4.1 shows examples of this interval for values of β and γ used in the trial calculations. The relationship between β and γ and the mean, mode and standard deviation of the various prior distributions, is displayed in Appendix D.

In general, the interval for θ_0 determines a prior distribution of the observed ability that is not a probability density function. A proper probability density function could always be obtained, however, by normalizing the constricted distribution - graphically, this meant changing the vertical scale (represented by κ in equation (3.8)) until the area underneath the curve of the prior distribution was equal to one.

It is worth noting here that a prior (beta) distribution with parameters a and b , functions within a beta-binomial analysis as though it were supplying $a + b$ items to the test (Novick and Jackson, 1974:126). These 'extra' items are, in a sense, mathematical equivalents of the information contained within the prior. In general, a prior which supplies a large amount of information on the most probable location of the examinee's observed (or true) ability, tends to be more peaked than one which does not supply nearly as much information of this kind. The graphs of distributions shown in

Table 4.1 θ_0 Intervals for Stated Values of β and γ

Parameters		Interval for θ_0
β	γ	
0	0	$[0,1]$
0	0.05	$[0,0.95]$
0	0.20	$[0,0.80]$
0.25	0	$[0.25,1]$
0.25	0.05	$[0.25,0.96]$
0.25	0.20	$[0.25,0.85]$

Figure 4.1 (including those distributions with 'infinite' peaks) clearly reveal this relationship.

Posterior Probabilities

Since posterior probabilities $\Pr(\theta_0 \geq \theta_c | x)$ were important elements in the calculation of cutting scores, they were examined in terms of the effects of test length, true criterion level, guessing and carelessness parameters and prior distribution. Appendices E1 - E6 report the results.

What was the effect of test length on the posterior probability? The results in Appendices E1 and E4 indicate that for a given combination of elements, the posterior probability for a given raw proportion score (x/n) on the ten item test was always greater than the corresponding posterior probability on the five item test. For example, with the uniform distribution $B(1,1)$, and with $(B,\gamma) = (0.25,0.20)$ and $\theta_c = 0.80$, the posterior probability for $x/n = 5/5 = 1$ on the five item test was 0.60, while the corresponding posterior probability for $x/n = 10/10 = 1$ on the ten item test was 0.81. It appeared that it would be possible to obtain larger maximum posterior probabilities simply by increasing the length of the test. This conclusion was not unexpected since, in practice, a ten item test would provide relative to a five item test, the greater amount of (sample) information on the most likely location of the examinees ability in the continuum.

Within tests prior distributions affected posterior probabilities in a rather more complicated way. On the one hand, when the mode of the prior distribution was made greater than the set criterion level the posterior probability was observed to increase; on the other hand, when the mode was made less than the set criterion level the probability was observed to decrease. For example, in the five item test the posterior probabilities at the 0.80 true criterion level were 0.80 and 0.56 for the $B(8,2)$ and $B(3,2)$ prior distributions respectively. Figure 4.1 shows that the mode of the former prior was above 0.7 whereas for the latter it was below this value. Similar results applied to the ten item test.

These observations, if applied to a group of examinees, would suggest that examinees who might initially be characterized as better performers (mode of their prior being greater than the set

criterion level), would be expected to obtain the greater posterior probabilities.

Moreover, the results also showed that a sufficient increase in the length of the test would always lead to a decrease in the difference between posterior probabilities based on different priors. This means that results derived from the Bayesian methodology will always tend to agree with the results derived with classical sample methods, when the weight of information supplied by test data is greater than that supplied by the prior distribution.

The parameters β and γ had opposite effects on the posterior probability. This probability was observed to increase with increasing values of γ when $\beta = 0$. Interpreted substantively, this would mean that if an examinee were careless the measurement model would credit him with 'extra probability' to compensate him for careless errors made in answering the test items. On the other hand, the posterior probability was observed to decrease with increasing values of β when $\gamma = 0$. In this case the measurement model would have penalized him for guessing.

Cutting Scores

Appendices F1 and F2 present cutting scores for the five and ten item tests, for certain combination of elements listed on page 34. Cutting scores increased when the cost of classifying a master as a non-master was decreased with respect to the cost of classifying a non-master as a master. That is, the cutting scores were observed to increase when the loss ratio C_s / C_f was decreased. This effect occurred for all combinations of β and γ and criterion levels, and was independent of test length. (Only results for the 0.70 and 0.80 true criterion levels have been tabulated.)

For some combination of elements it was not possible to determine a cutting score whose posterior probability was close to the corresponding loss ratio (see equation 3.4). In any case, at the cutting score the difference between the posterior probability and this ratio was not more than 0.10.

The individual effects of β and γ on the cutting score were the reverse of the individual effects of these parameters on the

posterior probability. When β was kept constant and γ was increased the cutting score was observed to decrease for fixed loss ratio and criterion level. Thus for the ten item test ($B(3,2)$, $\beta = 0$, $C_s / C_f = 5 / 1$, and $\bar{\theta}_c = 0.70$), the cutting score was 6, 5, and 3, when γ was 0, 0.05, and 0.20, respectively. In contrast, the cutting score was observed to increase for greater values of β when γ was held constant. Thus for the ten item test ($B(3,2)$, $\gamma = 0$, $C_s / C_f = 5 / 1$, and $\bar{\theta} = 0.70$), the cutting score was 6 and 7 when β was 0 and 0.25, respectively. These results clearly indicate the sensitivity of the cutting score to changes in the values of β and γ . In the examples just shown the cutting score took its smallest value 3 when $(\beta, \gamma) = (0, 0.20)$, and took its largest value 7 when $(\beta, \gamma) = (0.25, 0)$. In practice, a cutting score of 3 would result in more of the examinees being classified as masters than would be the case with a cutting score of 7. Clearly, then, mastery/non-mastery decisions in criterion-referenced testing (assuming the validity of the present model) could be markedly affected by the levels of examinee guessing and carelessness.

The shape of the prior distribution was instrumental in determining the possible range of values of the cutting score, for a test of fixed length. In general, the cutting score tended to be less for priors that had most of their bulk at the top end of the ability scale, but greater for priors that had most of their bulk at the bottom end of the ability scale. In some cases the test had too few items to provide meaningful (sample) information for calculating a non-zero cutting score; that is, the likelihood was completely dominated by the prior distribution.

Statistics of Posterior Distributions

Appendices G1 and G2 show the mean, mode, and standard deviation of each posterior distribution at the cutting scores. Without exception the mean and mode tended to increase, but the standard deviation tended to decrease as the loss ratio C_s / C_f was decreased. These statistics changed in value appreciably over the range of values assumed by the loss ratio. Also tabulated in these appendices are the parameters of the corresponding prior distributions.

CHAPTER 5

PRACTICAL APPLICATION OF THE MODEL: THE DETERMINATION OF CUTTING SCORES FOR CRITERION-REFERENCED TESTS FROM THE AUSTRALIAN STUDIES IN SCHOOL PERFORMANCE

The Australian Studies in School Performance (Bourke and Keeves, 1977) investigated by survey the standards of performance in the basic skills of literacy and numeracy of 10- and 14-year-old students in Australian schools. Stratified random samples of students were selected with a probability proportional to school size, from primary and secondary schools in the Australian States and Territories, yielding data for 6628 10-year-old and 6247 14-year-old students when testing had been completed.

Tests and Sub-Tests of Literacy and Numeracy

For each of the two age levels two criterion-referenced tests were used - a Reading Test and a Numeration Test (ACER, 1976a; ACER, 1976b). Each test was subdivided into a number of sub-tests varying in length and degree of content homogeneity. Appendix B shows how the items were distributed among the sub-tests.

Items in the tests were constructed so that they each related to a specific task or skill concerning an important facet of basic literacy or numeracy. Indeed, at the time of their construction, the nature and composition of the tests reflected the test-developers' beliefs of what might reasonably be considered as essential areas of content, ability, and usage, that needed to be surveyed so that worthwhile recommendations could be made concerning the percentages of students who had mastered these basic skills.

- 1 Reading Tests. These consisted of 29 items for the 10-year-olds and 33 items for the 14-year-olds, all but one item being of the multiple-choice type. The skills surveyed by these tests included: basic knowledge of the alphabet, reading simple sentences (referred to as the picture items), basic linguistic competence, reading comprehension of continuous prose, and use of a newspaper. Eleven sub-tests were formed for each age level from these tests.

- 2 Numeration Tests. These consisted of 33 items for each age level, most items being of the constructed-response type. The skills surveyed by these tests included: counting, recalling and applying the fundamental arithmetical operations, solving simple money problems, reading simple graphs and tables, and being able to read common measuring instruments. Twelve sub-tests were formed for the 10-year-olds and thirteen for the 14-year-olds.

Since a primary objective was the calculation of percentages of students who had not mastered the basic skills related to each of the sub-tests, it was necessary to first calculate a cutting score for each sub-test. With this information it then was a simple task to determine the percentages of students whose observed scores on each sub-test and total test were less than the cutting scores. Students whose scores fell below the cutting score for the test or sub-test were classified as non-masters of the domain of tasks referenced by that test or sub-test.

In order to determine the best course to follow in the calculation of cutting scores on tests from the Australian Studies in School Performance, preliminary calculations were carried out on the following three sub-tests.

- 1 Picture Items Sub-Test. These contained five multiple-choice items with four alternatives per item. Each item consists of a simple sentence, and students were asked to identify which scene was described by the sentence.
- 2 Linguistic Competence Sub-Test. These contained seven multiple-choice items with three alternatives per item. For each item in the sub-test, students were required to choose an appropriate alternative to complete a given sentence.
- 3 Operations on Numbers Sub-Test. These contained eight open-ended items to which students supplied written answers. This sub-test assessed students' knowledge of basic arithmetical operations on one and two digit whole numbers.

Each sub-test was considered to be a fairly good approximation of a random sample of items from the homogeneous item population underlying that sub-test. The same sub-tests were used at each age level, allowing

valid direct comparisons to be made among the observed abilities of students from the two age levels.

Specifying Prior Distributions for Tests and Sub-Tests

Table 5.1 shows posterior probabilities for the three sub-tests based on the uniform prior distribution. The use of the uniform prior distribution to represent situations where no significant prior knowledge exists is known as Bayes' postulate. Its use for this purpose has been controversial, but this has not prevented it being used as a first approximation. Box and Tiao (1973:25) have provided a good discussion on problems associated with the specification of noninformative prior distributions.

A drawback to using the uniform prior in a beta-binomial analysis of test response data is that it operates as though two items had been added to the sample test data. That is, the test length is effectively increased by two items (see page 36), and this may be of some concern when short tests are being used. Thus the posterior probabilities reported in Table 5.1 do not strictly refer to the no knowledge situation, but this was not a concern for the present study since the intention was to derive cutting scores from informative priors anyway. Indeed, the main function of the uniform prior was to provide a standard for comparing the magnitude of effects arising from informative priors.

Posterior probabilities recorded in Table 5.1 were calculated assuming the examinees guess randomly and that each examinee's level of carelessness could be represented by the value $\gamma = 0.05$. Calculations based on random guessing are at best a rough approximation, and the value 0.05 for the level of carelessness is a subjective choice for which there well might be another reasonable value. Nevertheless the values chosen for the parameters were viewed to be reasonable for making inferences from the data of the Australian Studies in School Performance.

For each score on the three sub-tests Table 5.1 shows the posterior probabilities derived from the 'basic' model ($\beta = 0$, $\gamma = 0$; first row) and those derived from the 'corrected' model ($\beta = 0$, 0.25, 0.33, and $\gamma = 0.05$; second row). On the one hand, where the test consists of completion type items, as in the sub-test - Operation on Numbers, the posterior probabilities derived from the

Table 5.1 Posterior Probability that Examinee's Observed Ability is Greater than $\bar{\theta}_c$ given a Uniform Prior Distribution^a

Sub-test	Length of test	Score on test	Posterior distribution	True criterion level $\bar{\theta}_c$					
				0.70	0.75	0.80	0.85	0.90	0.95
Operations on Numbers	8	8 ^a	B(9,1)	96 ^b	92	87	77	61	*
				97	92	87	77	63	*
	7	7	B(8,2)	80	70	56	40	*	*
				86	77	65	49	*	*
	6	6	B(7,3)	54	40	*	*	*	*
				64	51	*	*	*	*
Picture Items	5	5	B(6,1)	88	82	74	62	47	*
				77	70	61	49	*	*
	4	4	B(5,2)	58	47	*	*	*	*
				46	*	*	*	*	*
Linguistic Competence	7	7	B(8,1)	94	90	83	73	57	*
				86	81	72	58	46	*
	6	6	B(7,2)	74	63	50	*	*	*
				61	53	42	*	*	*
	5	5	B(6,3)	45	*	*	*	*	*
				*	*	*	*	*	*

^a For each score on the test the probabilities entered in the first row are derived from the basic model; probabilities entered in the second row are derived from the revised model $\theta_0 = \beta + (1 - \beta)(1 - \gamma)$. All probabilities are entered as percentages, $\gamma = 0.05$ for all sub-tests, $\beta = 0.25$ for the Picture Items sub-test, and $\beta = 0.33$ for the Linguistic Competence sub-test.

^b Probabilities estimated from tables of the incomplete beta function (Pearson, 1968).

* Indicates $\Pr(\theta_0 \geq \theta_c | x) < 0.40$.

corrected model were consistently greater than those derived from the basic model. Indeed, the difference between the two sets of probabilities was marked at lower scores. This observation is readily explained by noting that at lower score levels the carelessness factor became more significant and as a result produced an inflation in the observed ability which in turn lead to an inflation in the probabilities. On the other hand, for the other two sub-tests, which contain multiple-choice items ($\beta = 0.25$ or 0.33), the guessing factor consistently dominated the effects due to the carelessness factor ($\gamma = 0.05$), with the result that posterior probabilities were reduced in magnitude in comparison with those derived from the basic model.

In most instances of practical research, the examiner does have prior knowledge, or at least an expectation, of an examinee's ability in relation to the objectives assessed by the test. The simplest evidence of this is in the defining of a criterion level for the test. It is often expressed in terms of an expected average ability of the group of examinees, or the likely ability level of one examinee. Hence a distribution with its mean or modal value set near the criterion level could serve as a suitable prior of ability. In fact many sources of information usually exist in one form or another from which a useful prior may be constructed (Winkler, 1967).

In the present study prior knowledge of each examinee's ability was identified with the performances of all students at the appropriate age level on the reading test or numeration test or sub-test under consideration. These observed score distributions of the performances of students (Appendices A1 - A4) were translated into informative prior distributions which have been summarized in Table 5.2. The table shows that each prior had been characterized by a beta function with appropriately determined parameters. The parameters were estimated using the fractile assessment procedure (Novick and Jackson, 1974:167) which fits a function from the beta class to the observed raw observations. The degree of fit of a specified prior distribution was assessed by calculating its mean and mode and comparing these values with the corresponding empirical values. If the degree of fit was assessed to be poor the estimated parameters were adjusted to conform with the empirical values of the statistics. However, this is an

Table 5.2 Selected Informative Beta Prior Distributions for Tests and Sub-Tests

Origin of prior	Age level	Prior distribution	Effective prior sample size	Theoretical distribution ^c			Empirical distribution	
				Mean	Mode	SD	Mean	Mode
Reading ^a Test	10	B(8,2)	10.0	0.800	0.875	0.121	0.786	0.862
	14	B(11,2.5)	13.5	0.815	0.870	0.102	0.823	0.889
Numeration ^a Test	10	B(7,1.5)	8.5	0.824	0.923	0.124	0.810	0.909
	14	B(5.5,1)	6.5	0.846	1.000	0.134	0.820	0.970
Operations on Numbers ^b	10	B(2.5,1)	3.5	0.714	1.000	0.231	0.762	1.000
	14	B(8.5,1)	9.5	0.895	1.000	0.095	0.918	1.000
Picture Items ^b Sub-Test	10	B(17,2)	19.0	0.895	0.941	0.069	0.884	1.000
	14	B(50,1)	51.0	0.980	1.000	0.019	0.972	1.000
Linguistic ^b Sub-Test	10	B(9.5,1)	10.5	0.905	1.000	0.087	0.918	1.000
	14	B(24,1.5)	25.5	0.941	0.979	0.046	0.973	1.000

^a These refer to the total test.

^b These refer to sub-tests.

^c Values of statistics obtained using tables in Pearson (1968).

approximate procedure for specifying a prior, for when the observed distribution is highly skewed the agreement between the estimated beta distribution and the observed distribution may not be perfect in any sense of the word. It was noted on page 36 that the use of a prior distribution with parameters a and b is mathematically equivalent to adding $a + b$ items to the test. Adding information from the prior to the sample test data enabled the Australian Studies in School Performance to use short sub-tests. Table 5.2 shows that the $a + b$ values ranged from 3.5 to 51.

Cutting Scores for Tests and Sub-Tests

Table 5.3 reports cutting scores based on the corrected model, for the three sub-tests. Three prior distributions were tried at each age level and for each test. They were a uniform prior, a prior based on the scores of all students on either the total reading or numeration tests, and a prior based on the observed scores of all scores on the particular sub-test. Again, calculations were made for six criterion levels ranging in value from 0.70 to 0.95.

Cutting scores were estimated using equations (3.4), (3.10), and (3.11). Moreover, the ratio C_f / C_s was assumed to be approximately unity. This assumption asserted that making a false positive error was no better nor worse than making a false negative error. In addition, the posterior probability that the observed ability exceeded the set criterion level was chosen to be approximately equal to 0.5, taking into consideration the necessity for integral cutting scores. However, the value of the probability was not allowed to drop below 0.45 when defining the cutting scores.

Corresponding to each posterior distribution the cutting score was observed to increase in magnitude as the criterion level was increased. A higher criterion of performance demanded a more stringent passing score on the test. In the examples presented in Table 5.3 the cutting scores derived from the informative priors were generally smaller than or equal to those derived from the uniform prior. This is because the informative priors were based on all students' observed score distributions which were highly skewed, with modes near the top end of the ability scale. Moreover, prior distributions based on collateral information from the performances of all students on the

Table 5.3 Estimated Cutting Scores^a for Tests Based on Three Different Prior Distributions

Sub-test	Age level	Length of test	Posterior distribution	True criterion level $\hat{\theta}_c$					
				0.70	0.75	0.80	0.85	0.90	0.95
Operations on Numbers	10	8	B(1+x,9-x) ^b	6	6	7	7	8	*
			B(7+x,9.5-x)	4	5	6	6	7	8
			B(2.5+x,9-x)	5	6	6	7	8	*
	14	8	B(1+x,9-x)	6	6	7	7	8	*
			B(5.5+x,9-x)	4	5	6	6	7	8
			B(8.5+x,9-x)	3	4	5	6	7	8
Picture Items	10	5	B(1+x,6-x)	4	5	5	5	*	*
			B(8+x,7-x)	3	4	4	5	5	*
			B(17+x,7-x)	1	2	3	4	5	*
	14	5	B(1+x,6-x)	4	5	5	5	*	*
			B(11+x,7.5-x)	2	4	4	5	*	*
			B(50 ^c ,6-x)	0	0	0	0	0	3
Linguistic Competence	10	7	B(1+x,8-x)	6	6	7	7	7	*
			B(8+x,9-x)	5	6	6	7	7	*
			B(9.5+x,8-x)	4	5	5	6	6	*
	14	7	B(1+x,8-x)	6	6	7	7	7	*
			B(11+x,9.5-x)	5	6	6	7	*	*
			B(24+x,9.5-x)	1	2	3	4	5	7

^a Estimated cutting scores reflect situation where $\Pr(\theta_0 \geq c \mid \text{data}) = 0.50$ but in any case the probability is not less than 0.45, and $C_f = C_s$.

^b Beta distribution where x is the score on the test. For each test the first entry incorporates a uniform prior, the second the prior based on performance of all subjects on the total reading or numeration test, and the last the prior corresponding to performance of all subjects on the particular sub-test under investigation.

^c To ease the effort of calculation the first parameter, a , for this posterior distribution was assumed equal to 50 for all values of x .

* Indicates $\Pr(\theta_0 \geq \hat{\theta}_c \mid x) < 0.45$.

sub-tests gave the consistently lowest cutting scores.

The superior performance on the 14-year-old students compared to the performance of 10-year-old students was reflected in their lower cutting (or passing) scores. Table 5.3 shows that cutting scores for the 14-year-olds were often less than the corresponding cutting scores for the 10-year-olds. These results were at first perplexing and seemed to be contra-intuitive. Ordinarily one might have reasonably expected that because of the superior performance of the 14-year-olds their passing scores should have been much higher than those for the 10-year-olds. However, this apparent paradox was illusory since in terms of the Bayesian statistical analysis cutting scores are determined in large measure by the shape of the prior distributions. Indeed, for a given test length, the more the bulk of the prior is displaced toward the top end of the ability scale the lower will be the cutting score, other things being equal. This was vividly illustrated by the $B(50,1)$ prior used with the Picture Items Sub-Test. In that case the prior swamped the likelihood, yielding a zero cutting score up to the 0.90 criterion level.

These preliminary calculations with the three sub-tests suggested that the following conditions were to be followed when calculating cutting scores for all tests and sub-tests in the Australian Studies in School Performance:

- 1 A true criterion level of 0.80;
- 2 β values equal to the weighted average probability of randomly guessing the correct answer for items in the tests and sub-tests;
- 3 $\gamma = 0.05$;
- 4 Prior distributions based on the distribution of students' observed scores, at each age level, on either the whole Reading Test or the whole Numeration Test;
- 5 Loss ratios as close to 0.50 as possible, yet yielding integral cutting scores. And in no case were posterior probabilities at a given cutting score to be less than 0.40.

Bourke and Keeves (1977:Chapter Three) have discussed these assumptions, and pointed out their implications for the findings of the Australian Studies in School Performance.

Table 5.4 Cutting Scores (c) for the Reading and Numeration Sub-Tests and Whole Tests

Test	10-year-old level				14-year-old level			
	n	β	c	Pr ^d	n	β	c	Pr
<u>Reading Sub-Tests</u>								
Picture Items	5	0.25 ^b	4	46 ^d	5	0.25	4	54
Linguistic Competence	7	0.33	6	47	7	0.33	6	54
Paragraph Reading	4	0.25	3	41	6	0.25	4	58
Newspaper Comprehension	5	0.25	4	46	9	0.25	7	47
Newspaper Information	7	0.25	6	56	5	0.25	4	54
Social (Newspaper Total)	12	0.25	10	54	14	0.25	11	46
Classroom	17	0.28	14	51	19	0.28	15	44
Words in Context	16	0.29	13	42	18	0.28	15	55
Use of Reference Material	7	0.25	6	60	5	0.25	4	58
Comprehending Prose	6	0.25	5	54	10	0.25	8	53
Anchor Items ^Q	16	0.29	13	48	16	0.29	13	53
<u>Whole Test</u>	29	0.27	24	54	33	0.27	27	54
<u>Numeration Sub-Tests</u>								
Numbers (Whole and Fraction)	18	0.01	13	46	-	-	-	-
Whole Numbers	-	-	-	-	11	0.02	8	60
Fractions	-	-	-	-	6	0.07	4	47
Measurement	10	0.06	7	48	11	0.07	8	56
Money	5	0	3	52	5	0.04	3	44
Recall and Manipulation	16	0.04	12	56	12	0.07	9	59
Application	17	0.05	12	41	21	0.04	16	55
Social	19	0.04	14	50	16	0.04	12	57
Classroom	14	0.04	10	47	17	0.06	12	42
Addition	5	0	3	52	4	0.05	2	37
Subtraction	4	0	2	41	6	0.03	4	50
Multiplication	4	0	2	41	6	0.06	4	50
Division	4	0	2	41	7	0.03	5	57
Anchor Items ^c	8	0	6	41	8	0	5	41
<u>Whole Test</u>	33	0.05	25	53	33	0.04	25	54

^a $\Pr(\theta_0 \geq \theta_0 \mid c)$ based on whole test priors.

^b Values represent the weighted average probability of randomly guessing the correct answer for items in the tests.

^c Items in this sub-test are identical for both age levels.

^d Decimal points omitted.

Table 5.4 shows cutting scores for the Reading and Numeration sub-tests and whole tests that were derived using the above conditions.

CHAPTER 6

CONCLUSION

This study has developed a two-parameter linear probabilistic model of criterion-referenced measurement, which relates an examinee's observed ability to his true ability through a guessing parameter and a carelessness parameter both of which are probabilities. In the model, the true ability has been conceptualized as the proportion of items in the hypothetical population of items that the examinee could answer correctly were he given the opportunity to do so. It is a latent variable that cannot be observed directly, but must be inferred from the person's test score. The observed ability, which is also a latent variable, is conceptualized as the proportion of items in the hypothetical population of items that the examinee could answer correctly given that he would guess and be careless as described by the model when responding to the test items. Thus the true ability reflects an ideal testing situation uncontaminated by extraneous factors that may produce systematic errors in the test scores, while the observed ability reflects testing situations where guessing and carelessness are operative and produce systematic errors in the test scores. Consequently, for example, it is possible that when the examinee responds to different samples of items from the one population of items, the degree of guessing and carelessness may change sufficiently to produce variation in the estimated observed ability across the samples. However, his true ability would, for each sample of items, remain necessarily constant at a value representing his performance for the whole population of items.

Although a more general criterion-referenced measurement model could have been specified in this study, with additional parameters above those for guessing and carelessness, this was not done because it would have been quite difficult to justify the need for the extra parameters within the survey context of the Australian Studies of School Performance, let alone justify in the fixed parameter format the values that might have been posited for them. Certainly the addition of extra parameters would have entailed working with a less parsimonious and mathematically more complex model.

The main reason for developing a criterion-referenced measurement model with corrections for guessing and carelessness was to assist with the interpretation of test data obtained by the Australian Studies in School Performance project. In that project a primary objective was to survey 10- and 14-year-old Australian students to estimate the proportions of students who had been able to master during the course of their lives a variety of basic skills that would indicate their levels of numeracy and literacy. The estimates were derived by utilizing a Bayesian inferential procedure which incorporated the developed measurement model within a beta-binomial, decision-theoretic framework based on a two state conception of mastery and non-mastery. For each test and sub-test of the project a cut-off score was determined on the raw score scale corresponding to a 0.80 criterion level (standard of performance) for the particular population of items; students whose raw scores exceeded or equalled a cut-off score were assumed to have mastered the basic skills that were assessed, and students whose raw scores were less than the cut-off score were considered not to have mastered those skills. The proportions were estimated by counting the numbers of students who were classified as masters. In deriving the cut-off scores the decision-theoretic part of the Bayesian inferential procedure was structured so that the cost of misclassifying a master as a non-master was approximately equal to the cost of misclassifying a non-master as a master.

The application of this model in a Bayesian framework, to test data from the Australian Studies in School Performance Project has, it is believed, yielded more reliable estimates of the proportions of students who had mastered the basic skills of numeracy and literacy than could have been possible within a classical statistical framework. In applying this model the parameters for guessing and carelessness were assigned constant values a priori. These values were considered to be 'reasonable' descriptors of the various testing situations, given the purposes of the testings and the mastery/non-mastery nature of the inferences to be made. In that project, average constant values were chosen for two reasons. First it would have been impracticable to assign individual values because of the logistics that would have been involved. Second the intention was to estimate the overall performances of groups and subgroups of 10- and 14-year-old students and not the performances of individual students. For this reason, all

students were presumed to guess randomly with the same probability of chance success, and all students were also presumed to be equally careless with the same probability of chance carelessness.

The results of a computer simulation of the model strongly suggest that biasing effects, such as those arising from the response sets of guessing and carelessness, can markedly affect posterior probabilities which in turn can affect cutting scores derived from them. It is highly probable that when such response sets are operating during testing, and that if they are not taken into consideration when calculating posterior probabilities and cutting scores, wrong mastery decisions could be made. Indeed it would seem to be good practice for the examiner to first make a preliminary survey of his students and testing environment to determine the probable extent of bias in the raw scores due to the influence of response sets and other sources of systematic error. If certain of these sources are identified as possible significant contributors of bias it would pay the examiner to correct for them in order to increase the accuracy of his mastery decisions. Unfortunately, most applications of the Bayesian decision-theoretic procedure discussed in the literature have failed to consider the presence of systematic error when calculating posterior probabilities and cutting scores, and when making mastery decisions. In criterion-referenced testing where absolute measures are made of examinees' performances it is essential that corrections be made for the effects of significant biasing errors.

The two parameter model that has been developed in this study is an appreciable step in this direction. It enables the examiner to explore the consequences of choosing certain values of the guessing and carelessness parameters, and allows him to select those values of the parameters which he feels are congruent with the actual reality of the testing situation. The model does allow the decision-maker to make more accurate mastery decisions about students' true abilities than might be possible with a model that fails to consider the effects of systematic error arising from guessing and carelessness, in situations where these response sets are operative.

Further research is needed to test the efficacy of assigning probability distributions to the parameters instead of keeping them fixed at particular values as has been done here. It is highly probable

that this research would reveal methods which would allow the deployment of a more flexible measurement model that would certainly have, at the least, greater logical validity. Indeed it is unlikely that in a practical testing situation the examiner would know every examinee's levels of guessing and carelessness to the same degree of accuracy or certainty. From the examiner's point of view it would be more realistic to specify a measurement model which could utilize interval estimates to accommodate this uncertainty rather than point estimates which presume, unrealistically, that these levels are known with absolute certainty. In this model, the posterior probability density function might then be represented as

$$f(\theta_0, \beta, \gamma \mid x) \propto g(x \mid \theta_0, \beta, \gamma) h(\theta_0, \beta, \gamma),$$

where $g(x \mid \theta_0, \beta, \gamma)$ is the likelihood function and $h(\theta_0, \beta, \gamma)$ is the joint prior distribution of the observed ability θ_0 , the guessing parameter β , and the carelessness parameter γ .

In order to calculate cutting scores it would then be necessary to first calculate the marginal posterior density function of θ_0 by integrating over the intervals for β and γ . However, to perform the integrations a functional form would have to be selected for $h(\theta_0, \beta, \gamma)$, a choice that is not that apparent considering the multivariate nature of this function. In any case the particular choice would need to be at the very least justifiable empirically and, perhaps tractable mathematically.

APPENDIX A1

Distribution of Scores on Reading Sub-Tests and Whole
Test for 10-Year-Old Australian Students

Percent of Students (Weighted) ^a												
No. of Items Correct	Simple Sentences	Linguistic Competence	Paragraph Reading	Newspaper Comprehension	Newspaper Information	Social (Newspaper Total)	Classroom	Words in Context	Use of Reference Materials	Comprehending Prose	Anchor Items	Whole Test
0	0.3	0.4	2.3	6.5	0.7	0.5	0.0	0.1	0.3	1.1	0.0	0.0
1	0.3	0.5	9.8	14.8	1.1	0.6	0.1	0.1	1.0	4.4	0.0	0.0
2	0.6	1.3	22.3	21.6	2.9	1.3	0.1	0.1	1.9	9.1	0.1	0.0
3	1.5	1.8	31.2	24.8	7.2	2.8	0.1	0.2	5.2	14.8	0.1	0.0
4	8.6	2.8	34.1	21.6	14.4	4.5	0.2	0.2	13.0	21.3	0.2	0.0
5	88.6	6.0		10.4	28.2	7.0	0.2	0.4	30.0	24.9	0.3	0.2
6		16.9			33.2	9.6	0.4	0.6	35.6	24.2	0.3	0.1
7		70.2			11.9	14.4	0.5	0.8	12.7		0.8	0.1
8						15.3	0.8	0.0			0.8	0.2
9						16.5	1.1	2.1			1.2	0.2
10						14.8	1.3	3.2			2.0	0.5
11						9.0	2.4	5.6			3.4	0.5
12						3.3	3.3	10.3			6.8	0.6
13							5.7	18.1			14.3	0.7
14							11.1	23.8			33.5	1.0
15							18.0	23.1			26.1	1.3
16							27.0	10.3			9.8	1.9
17							27.3					2.7
18												3.3
19												4.2
20												5.2
21												6.4
22												7.7
23												10.2
24												11.2
25												12.0
26												11.3
27												9.8
28												6.3
29												1.8
No. of Items in Sub-Test	5	7	4	5	7	12	17	16	7	6	16	29

^a There are missing data amounting to 0.3% of students (weighted) or 22 students (unweighted) associated with each Sub-test and with the Whole Test. These students did not attempt any item in the Reading Test but did participate in other parts of the testing program. They do not include the 19 students who attempted the test but failed to obtain any correct answers.

APPENDIX A2

Distribution of Scores on Reading Sub-Tests and Whole Test for 14-Year-Old Australian Students

No. of Items Correct	Percent of Students (Weighted) ^a											
	Simple Sentences	Linguistic Competence	Paragraph Reading	Newspaper Comprehension	Newspaper Information	Social (Newspaper Total)	Classroom	Words in Context	Use of Reference Materials	Comprehending Prose	Anchor Items	Whole Test
0	0.1	0.1	1.0	0.6	1.0	0.3	0.1	0.0	0.2	0.2	0.0	0.0
1	0.1	0.1	3.5	1.0	2.8	0.2	0.1	0.0	1.9	0.6	0.0	0.0
2	0.1	0.2	7.3	2.3	5.2	0.4	0.1	0.0	6.9	1.7	0.0	0.0
3	0.3	0.3	12.7	3.1	21.2	1.1	0.1	0.1	22.0	3.0	0.0	0.0
4	1.9	0.4	20.1	5.3	37.0	1.4	0.1	0.0	36.4	5.1	0.0	0.0
5	97.3	1.9	29.1	8.6	32.5	2.1	0.1	0.0	32.4	7.3	0.0	0.0
6		8.2	25.9	12.1		2.4	0.1	0.0		11.7	0.1	0.0
7		88.5		18.6		3.9	0.0	0.1		14.0	0.1	0.0
8				23.2		5.4	0.1	0.1		19.1	0.0	0.0
9				25.1		7.8	0.2	0.2		20.8	0.2	0.1
10						10.7	0.3	0.4		16.3	0.4	0.0
11						15.2	0.4	0.8			0.7	0.0
12						17.5	1.1	1.8			1.7	0.1
13						18.5	1.7	2.3			4.3	0.1
14						12.9	3.9	4.3			14.7	0.3
15							7.2	7.4			34.3	0.5
16							13.5	14.8			43.2	0.4
17							18.9	27.4				0.8
18							28.1	39.9				1.0
19							24.1					1.1
20												1.1
21												1.8
22												2.4
23												2.9
24												3.7
25												5.1
26												6.3
27												7.7
28												9.6
29												11.2
30												12.5
31												12.7
32												11.5
33												6.3
No. of Items in Sub-Test	5	7	6	9	5	14	19	17	5	10	16	33

^a There are missing data amounting to 0.3% of students (weighted) or 20 students (unweighted) associated with each Sub-test and with the Whole Test. These students did not attempt any item on the Reading Test but did participate in other parts of the testing program. They do not include the 10 students who attempted the test but failed to obtain any correct answers.

APPENDIX A3

Distribution of Scores on Numeration Sub-Tests and Whole Test for 10-Year-Old Australian Students

No. of Items Correct	Percent of Students (Weighted) ^a												
	Numbers	Measurement	Money	Recall/ Manipulation	Application	Social	Classroom	Addition	Subtraction	Multiplication	Division	Anchor Items	Whole Test
0	0.1	0.2	0.6	0.1	0.2	0.2	0.2	0.4	1.6	2.4	3.0	2.4	0.1
1	0.1	0.4	0.7	0.1	0.2	0.1	0.2	0.6	4.4	5.0	9.7	1.8	0.1
2	0.1	1.2	1.5	0.2	0.2	0.1	0.4	1.5	12.0	13.7	16.3	3.3	0.0
3	0.3	1.7	3.8	0.5	0.2	0.2	1.0	6.1	27.5	30.9	24.5	5.1	0.0
4	0.4	2.8	14.0	0.6	0.4	0.3	1.2	18.3	54.3	47.8	46.2	6.8	0.1
5	0.6	5.4	79.1	1.1	0.6	0.4	2.4	72.9				10.3	0.2
6	0.9	9.9		1.8	0.4	0.6	3.1					15.2	0.1
7	1.6	13.8		2.5	0.8	0.3	5.0					22.2	0.1
8	2.3	20.8		4.0	1.2	0.9	4.7					32.6	0.2
9	2.9	28.5		3.8	1.8	1.1	7.1						0.2
10	3.6	14.9		5.3	2.9	1.3	8.9						0.3
11	4.3			7.0	4.3	2.2	12.0						0.5
12	6.3			8.8	7.4	3.2	16.1						0.3
13	7.3			12.5	10.4	4.8	18.8						0.3
14	10.6			15.4	17.4	6.8	18.7						0.6
15	13.6			17.8	21.6	11.3							0.7
16	16.0			18.1	20.7	16.2							1.5
17	17.4				9.0	20.9							1.1
18	11.3					19.8							1.5
19						9.2							1.7
20													2.5
21													2.7
22													3.0
23													3.9
24													3.8
25													4.7
26													6.5
27													7.8
28													8.9
29													10.6
30													11.7
31													10.5
32													9.6
33													4.3
No. of Items in Sub-Test	18	10	5	16	17	19	14	5	4	4	4	8	33

^a There are missing data amounting to 0.2% of students (weighted) or 13 students (unweighted) associated with each Sub-test and with the Whole Test. These students did not attempt any item in the Numeration Test but did participate in other parts of the testing program. They do not include the 34 students who attempted the test but failed to obtain any correct answers.

APPENDIX A4

Distribution of Scores on Numeration Sub-Tests and Whole Test for 14-Year-Old Australian Students

Percent of Students (Weighted) ^a														
No. of Items Correct	Whole Numbers	Fractions	Measurement	Money	Recall/ Manipulation	Application	Social	Classroom	Addition	Subtraction	Multiplication	Division	Anchor Items	Whole Test
0	0.0	1.4	0.3	1.0	0.0	0.1	0.2	0.0	0.1	0.1	0.8	1.2	0.1	0.0
1	0.0	3.4	0.6	2.9	0.1	0.2	0.3	0.0	1.2	0.9	1.7	2.5	0.1	0.0
2	0.4	6.2	0.8	7.3	0.3	0.4	0.6	0.1	9.5	2.1	3.7	3.8	0.5	0.0
3	0.5	9.9	0.9	17.3	0.5	0.4	1.0	0.2	29.1	5.4	7.6	7.0	0.8	0.0
4	0.9	14.9	2.0	31.9	0.9	0.6	0.8	0.3	60.0	10.5	14.6	10.3	1.8	0.2
5	1.8	22.6	2.9	39.4	1.4	0.8	0.9	0.7		24.7	28.8	15.7	3.1	0.1
6	2.4	41.4	5.4		2.3	0.5	1.3	0.6		56.1	42.8	25.3	7.5	0.2
7	5.1		7.2		4.1	0.9	2.0	0.8				34.0	24.5	0.2
8	9.7		12.1		7.8	1.1	3.0	1.9					61.5	0.5
9	16.5		16.1		13.3	1.2	3.5	2.1						0.3
10	26.6		24.1		17.3	1.8	5.0	3.3						0.4
11	35.8		27.3		23.3	2.1	6.5	5.4						0.5
12					28.6	2.7	8.1	6.8						0.4
13						3.9	11.1	9.4						0.7
14						5.0	16.4	12.5						0.3
15						5.6	19.3	14.5						0.8
16						5.9	19.9	19.8						0.7
17						8.9		21.3						1.3
18						11.0								1.3
19						15.2								1.5
20						16.7								2.3
21						15.0								2.8
22														2.8
23														3.7
24														4.3
25														3.9
26														5.3
27														5.1
28														7.9
29														8.6
30														9.7
31														12.0
32														12.3
33														9.6
No. of Items in Sub-Test	11	6	11	5	12	21	16	17	4	6	6	7	8	33

^a There are missing data amounting to 0.2% of students or 16 students (unweighted) associated with each Sub-Test and with the Whole Test. These students did not attempt any item in the Numeration Test but did participate in other parts of the testing program. They should not be confused with an additional 14 students who attempted the Test but failed to obtain any correct answers.

APPENDIX B

Sub-Tests of the Whole Tests

(a) 10-Year-Old Level

Test	Items	n
<u>Reading Sub-Tests</u>		
Picture Items	2 - 6	5
Linguistic Competence	7 - 13	7
Paragraph Reading	14 - 17	4
Newspaper Comprehension	23, 24, 26 - 28	5
Newspaper Information	18 - 22, 25, 29	7
Social (Newspaper Total)	18 - 29	12
Classroom	1 - 17	17
Words in Context	2 - 13, 23, 24, 27, 29	16
Use of Reference Material	1, 18 - 22, 25	7
Comprehending Prose	14 - 17, 26, 28	6
Anchor Items	1 - 13, 19, 21, 27	16
<u>Whole Test</u>		29
<u>Numeration Sub-Tests</u>		
Numbers (Whole and Fraction)	1 - 3, 6, 8, 10 - 19, 25, 30, 32	18
Whole Numbers		-
Fractions		-
Measurement	4, 5, 7, 9, 24, 26, 28, 29, 31, 33	10
Money	20 - 23, 27	5
Recall and Manipulation	2, 3, 6, 8, 10 - 18, 24, 29, 32	16
Application	1, 4, 5, 7, 9, 19 - 23, 25 - 28, 30, 31, 33	17
Social	1, 4 - 6, 8, 14, 19 - 27, 29 - 31, 33	19
Classroom	2, 3, 7, 9 - 13, 15 - 18, 28, 32	14
Addition	10, 15, 19, 20, 22	5
Subtraction	11, 16, 21, 30	4
Multiplication	12, 17, 25, 27	4
Division	13, 18, 23, 31	4
Anchor Items	10 - 13, 15 - 18	8
<u>Whole Test</u>		33

APPENDIX B (continued)

(b) 14-Year-Old Level

Test	Items	n
<u>Reading Sub-Tests</u>		
Picture Items	2 - 6	5
Linguistic Competence	7 - 13	7
Paragraph Reading	14 - 19	6
Newspaper Comprehension	22 - 30	9
Newspaper Information	20, 21, 31 - 33	5
Social (Newspaper Total)	20 - 33	14
Classroom	1 - 19	19
Words in Context	2 - 13, 26 - 31	18
Use of Reference Material	1, 20, 21, 32, 33	5
Comprehending Prose	14 - 19, 22 - 25	10
Anchor Items	1 - 13, 20, 21, 27	16
<u>Whole Test</u>		33
<u>Numeration Sub-Tests</u>		
Numbers (Whole and Fraction)		-
Whole Numbers	10 - 13, 15 - 18, 22, 25, 33	11
Fractions	3, 4, 20, 21, 24, 28	6
Measurement	1, 2, 6, 8, 9, 14, 19, 26, 27, 31, 32	11
Money	5, 7, 23, 29, 30	5
Recall and Manipulation	1, 4, 10 - 13, 15 - 18, 31, 33	12
Application	2, 3, 5 - 9, 14, 19 - 30, 32	21
Social	2, 3, 5, 7, 8, 16, 19 - 25, 28 - 30	16
Classroom	1, 4, 6, 9 - 18, 27, 31 - 33	17
Addition	2, 10, 15, 32	4
Subtraction	3, 7, 11, 16, 19, 23	6
Multiplication	1, 12, 17, 20, 24, 31	6
Division	5, 13, 18, 25, 26, 28, 33	7
Anchor Items	10 - 13, 15 - 18	8
<u>Whole Test</u>		33

APPENDIX C

Program BAYES for Calculating:

- (a) Cutting Scores;
- (b) Posterior Probabilities $\Pr(\theta_0 \geq \theta_c \mid x = c)$; and
- (c) Means, Modes, and Standard Deviations of Prior
Distributions $\rho(\theta_0)$, and Posterior Distributions
 $f(\theta_0 \mid x = c)$

Description

Program BAYES has been written in the FORTRAN language. It is listed on the following pages with a sample of output. In its present form it will run on the CYBER-73, and it may be easily adapted to run on other similar computers. BAYES consists of two separate routines: a main program; and a subroutine called MDBETA which is not listed here. Details regarding MDBETA can be found in the International Mathematical Statistical Libraries (1974).

The function of the main program is to input data, output results, and to calculate values of variables depending on the probability values generated by MDBETA. The main program has been structured so that calculations will be performed in an iterative manner, descending in order, over the set of values of:

- 1 test lengths;
- 2 prior distributions;
- 3 guessing and carelessness parameters;
- 4 criterion levels; and
- 5 loss ratios.

For example, for each prior distribution in a set, iterations will be made over the set of values of the guessing and carelessness parameters, and then in turn for each of the latter pair, iterations will be made over the set of criterion levels.

Subroutine MDBETA calculates the probability P that a random variable, whose probability density function is a beta distribution with parameters A and B , is less than or equal to X .

Cutting scores

BAYES determines a cutting score called CUT for each combination of criterion level and loss ratio, as follows:

For each score M lying in the interval from MIN to MAX (which are supplied by the examinee), BAYES examines the magnitude of

$$SEPT = \left| \frac{Q}{1 + Q} - PROB \right|$$

where Q is the loss ratio CSUCC / CFAIL, and PROB is one minus the posterior probability POSTP, at score M. If SEPT > 1 the CUT is taken to be MIN - 1, otherwise the magnitude of M is increased in steps of one score unit until the smallest value of SEPT is found corresponding to the defined cutting score. If M should reach the MAX score, CUT is put equal to MAX + 1. Obviously, the cutting score obtained by the user is dependent on the specified interval [MIN, MAX].

List of variables

NTEST:	Number of tests in the current job	}	for test J.
NO(J):	Number of items		
NPRIOR(J):	Number of prior distributions		
NGC(J):	Number of pairs of guessing and carelessness parameters		
NCRIT(J):	Number of criterion levels		
NLOS(J):	Number of loss ratios		
XMIN(J) and XMAX(J):	Minimum and maximum scores	}	for the I th prior distribution, for test J.
AA(I,J); BB(I,J):	Beta function parameters		
BETA(I,J); GAMMA(I,J):	Guessing and carelessness parameters		
CRIT(I,J):	Criterion levels		
CFAIL(I,J); CSUCC(I,J):	Losses		
PRME; PRMO; PRSIG:	Mean, mode, and standard deviation of prior distribution		
LIML(L); LIMU(L):	Lower and upper limits of the posterior probability integral, for the L th criterion level		

POSTP(N,L):	Posterior probability	} for the N th loss ratio and L th criterion level.
CUT(N,L):	Cutting score	
MEAN(N,L):	Mean of posterior	
MODE(N,L):	Mode of posterior	
SIG(N,L):	Standard deviation of posterior	

Input to BAYES

The information punched on each card in the data deck is described below by: type of input data; upper limit of number of cases of this data, if applicable; columns in which data is to be punched; and the format of the data.

Card No.	Information
1	Title card; col. 1 - 80; 10A8
2	Number of tests in current job; max. 10; col. 1 - 4; 14
3	NO, NPRIOR, NGC, NCRIT, NLOS, XMIN, XMAX; max. 10 for all variables; col. 1 - 28; 7I4 (one card is required for each test in the current job)
4	AA(1,1), BB(1,1), AA(2,1), BB(2,1),... ; max. 15 pairs; col. 1 - 80; 5(F8.2,F8.2)
5	BETA(1,1), GAMMA(1,1), BETA(2,1), GAMMA(2,1),... ; max. 25 pairs; col. 1 - 80; 5(F8.2,F8.2)
6	CRIT(1,1), CRIT(2,1),... ; max. 10; col. 1 - 80; 10F8.2
7	CFAIL(1,1), CSUCC(1,1), CFAIL(2,1), CSUCC(2,1),... ; max. 15 pairs; col. 1 - 80; 5(F8.2,F8.2)

Cards 4 - 7 are repeated for each test in the current job.

```

1      PROGRAM HAYES(INPUT,OUTPUT)
2      INTEGER XMIN(10),XMAX(10),CUT(15,10)
3      REAL TPR,CDF(10),CML(10),MEAN(15,10),MODE(15,10)
4      REAL SIG(15,10)
5      DIMENSION TITLE(10),NO(10),NPRIOR(10),NGC(10),NCRIT(10),NLOS(10)
6      DIMENSION AA(15,10),JR(15,10),PETA(25,10)
7      DIMENSION GAMMA(25,10),CRIT(10,10),CFAIL(15,10),CSUCC(15,10)
8      DIMENSION AX(35,1),BX(35,1),CONST(35),POSTP(35,10)
9      DIMENSION AS(15,10),BS(15,10)
10
11      C READ TITLE CARD
12      READ 2,(TITLE(I),I=1,10)
13      2 FORMAT(10A8)
14      C READ NUMBER OF TEST RUNS
15      READ 4,NTEST
16      4 FORMAT(I4)
17
18      C READ NUMBER OF INDEPENDENT PARAMETERS FOR EACH TEST
19      READ 6,(NO(J),NPRIOR(J),NGC(J),NCRIT(J),NLOS(J),XMIN(J),
20      1 XMAX(J),J=1,NTEST)
21      6 FORMAT(7I4)
22      C READ VALUES OF PARAMETERS - ONE SET FOR EACH TEST
23      DO 15 J=1,NTEST
24      IO=NPRIOR(J)
25      IP=NGC(J)
26      IN=NCRIT(J)
27      IR=NLOS(J)
28      READ 8,(AA(I,J),JR(I,J),I=1,IO)
29      READ 8,(PETA(I,J),GAMMA(I,J),I=1,IP)
30      READ 10,(CRIT(I,J),I=1,IN)
31      READ 8,(CFAIL(I,J),CSUCC(I,J),I=1,IR)
32      8 FORMAT(5(F8.2,F8.2))
33      10 FORMAT(10F8.2)
34      15 CONTINUE
35      PRINT 12,(TITLE(I),I=1,10)
36      12 FORMAT(1H1,6X,10A8)
37      C GET 4 LOOPS FOR CALCULATIONS
38      ICR=0.001

```

```

DO 100 J=1,NTEST
  MIN=XMIN(J)
  MAX=XMAX(J)
40  C NEXT LOOP IS OVER THE PRIORS
    IO=NPRIOR(J)
    DO 95 I=1,IO
      C NEXT LOOP IS OVER GUESSING AND CARELESSNES
45      IP=NGC(J)
        DO 90 K=1,IP
          IF (BETA(K,J).NE.0.0.OR.GAMMA(K,J).NE.0.0) GO TO 26
          IND=1
          XY=1.0
          XZ=0.0
          GO TO 27
26        XY=BETA(K,J)+(1.0-BETA(K,J))*(1.0-GAMMA(K,J))
          XZ=BETA(K,J)
55      27 CONTINUE
        C CALCULATE STATISTICS OF PRIOR
          A=AA(I,J)
          B=BB(I,J)
          X=XY
          CALL MUBETA(X,A,B,P,IER)
          P1=P
          X=XZ
          CALL MUBETA(X,A,B,P,IER)
          P2=P
          DO I=1.0/(P1-P2)
            A=AA(I,J)+1.0
            B=BB(I,J)
            X=XY
            CALL MUBETA(X,A,B,P,IER)
            P1=P
            X=XZ
            CALL MUBETA(X,A,B,P,IER)
70

```

```

      X2=0
      P1=CON*(P1-P2)*((A-1.0)/(A+1-1.0))
      AA=1.
      A=AY
      CALL MDXETA(X,A,,P,10)
      P1=P
      A=XZ
      CALL MDXETA(X,A,,P,10)
      P1=P
      TEMP=((A-2.0)*(A-1.0))/((A+3-2.0)*(A+1-1.0))
      PVAR=(CON*(P1-P2)*TEMP)-PVAR**2
      PRIG=SQRT(PVAR)
      IF((A+1).LT.4.) GO TO 113
      PRIG=(A-2.0)/(A+1-1.0)
      IF(PVAR.GT.XZ) PRIG=XZ
      IF(PVAR.GT.XY) PRIG=AY
      GO TO 112
110 PRIG=9.9
112 CONTINUE
2 NEXT LOOP IS OVER CRITERION VALUES
      L=NO. T(J)
      DO 35 L=1,L
      IF(IND.EQ.2) GO TO 29
28 XX=CRIT(L,J)
      LIMU(L)=1.0
      LIML(L)=XX
      GO TO 30
100 29 LIMU(L)=AY
      LIML(L)=BETA(K,J)+CRIT(L,J)*(1.0-BETA(K,J))*(1.0-GAMMA(K,J))
      30 CONTINUE
      DO 30 M=MIN,MAX
155 AX(M,1)=AA(I,J)+1
      A=AX(M,1)
      BX(M,1)=BB(I,J)+ND(J)-M
      B=BX(M,1)
      IF(IND.EQ.2) GO TO 31

```

```

110      X=XX
      CALL MDGETA(X,A,B,P,IER)
      CONST(M)=1.0
      POSTP(M,L)=1.0-P
      GO TO 80
115      *1 X=XY
      CALL MDGETA(X,A,B,P,IER)
      P1=P
      X=XZ
      CALL MDGETA(X,A,B,P,IER)
120      P2=P
      X=LIML(L)
      CALL MDGETA(X,A,B,P,IER)
      P3=P
      CONST(M)=1.0/(P1-P2)
125      POSTP(M,L)=CONST(M)*(P1-P3)
      GO CONTINUE

```

```

      C
      C CALCULATE CUTTING SCORES FOR ALL LOSS RATIOS AT THE PARTICULAR CRIT
130      IPOT=-1
      L=ENLOS(J)
      DO #2 N=1,IR
      Z=OSUCC(N,J)/CFAIL(N,J)
      RATIO=N/(1.0+Z)
135      DIFF=1.0
      DO #2 M=MIN,MAX
      IOO=M
      PRON=1.0-POSTP(M,L)
      SEPT=ABS(RATIO-PRON)
140      IF (SEPT.GE.DIFF) GO TO 33
      IF (1.EQ.MAX) IOO=M+1
      DIFF=SEPT
      #2 CONTINUE
145      #3 CUT(N,L)=IOO-1
      IF (CUT(N,L).NE.IPOT) GO TO 118
      MEAN(N,L)=MEAN(N-1,L)
      MODE(N,L)=MODE(N-1,L)

```

```

      SIG(N,L)=SIG(N-1,L)
      GO TO 122
118 IPOT=CUT(N,L)
2. CALCULATE STATISTICS OF POSTERIOR FOR EACH LOSS RATIO AND CRITERION
      AS(N,L)=AA(I,J)+CUT(N,L)
      A=AS(N,L)+1.0
115      BS(N,L)=BB(I,J)+BD(J)-CUT(I,L)
      B=BS(N,L)
      X=XY
      CALL MDGETA(X,A,B,P,IER)
      P1=P
116      X=XZ
      CALL MDGETA(X,A,B,P,IER)
      P2=P
      TEMP=(N,L)=CONST(IGC-1)*(P1-P2)*((A-1.0)/(A+3-1.0))
      A=A+1.0
117      X=XY
      CALL MDGETA(X,A,B,P,IER)
      P1=P
      X=XZ
      CALL MDGETA(X,A,B,P,IER)
      P2=P
170      TEMP=((A-2.0)*(A-1.0))/((A+3-2.0)*(A+3-1.0))
      TEMP2=(CONST(IGC-1)*(P1-P2)*TEMP)-M=AN(N,L)**2
      SIG(N,L)=SQRT(TEMP2)
      IF((A+3).GE.4.0) GO TO 123
175      MODE(N,L)=(A-3.0)/(A+3-4.0)
      IF(400E(N,L).LT.XZ) MODE(N,L)=XZ
      IF(400E(N,L).GT.XY) MODE(N,L)=XY
      GO TO 122
120 MODE(N,L)=9.9
119 CONTINUE
      82 CONTINUE
      85 CONTINUE

```

```

C PRINT THE RESULTS
185 PRINT #0, NO(J), AA(I, J), CB(I, J), BETA(K, J), GAMMA(K, J)
60 FORMAT(////90H POSTERIOR PROBABILITIES(1) CUTTING SCORES(2) A
100 STATISTICS OF DISTRIBUTIONS(3) 44EN//294 TEST L
LENGTH =, I2, //32H PRIOR = F(, F5.2, F5.2, 1H),
3//50H GUESSING PARAMETER = , F4.2,
190 4//50H CAPTESSON'S PARAMETER = , F4.2, //)
PRINT #1
61 FORMAT(55H (1) TABLE OF PROB(THETA GT CRITERION GIVEN SCORE)/
170H = CONST TIMES INTEGRAL OF POSTERIOR FROM LIM
21 TO L400//)
195 PRINT #2
62 FORMAT(120H -----)
1
PRINT #3, (CRIT(L, J), L=1, 10)
63 FORMAT(60H CRITERION VA
200 1L06/42H SCORE POSTERIOR CONST , 10(F4.2, 3X))
PRINT #2
DO 65 M=MIN, MAX
PRINT #4, M, 4X(, 1), 3X(M, 1), CONST(M), (POSTP(M, ,), L=1, 10)
64 FORMAT(11X, I2, 5X, 2H(, F5.2, 1H, F5.2, 1H), 3X, 11(F4.2, 3X))
205 CONTINUE
PRINT #2
PRINT #6, (LIMU(L), L=1, 10)
66 FORMAT(40H LIMITS OF THE LIMU =, 2X, 10(F4.2, 3X))
PRINT #7, (LIML(L), L=1, 10)
210 67 FORMAT(40H POSTERIOR INTEGRALS LIML =, 2X, 10(F4.2, 3X))
PRINT #2
PRINT #8

68 FORMAT(////33H (2) TABLE OF CUTTING SCORES//)
PRINT #2
215 PRINT #9, (CRIT(L, J), L=1, 10)
69 FORMAT(40H LOSS RATIO CRITERION VALUES/
100H 0=GOOD TO CFAIL , 10(2X, F4.2))
PRINT #2

```

```

220      DO 70 N=1,IR
          PRINT 71,CSUCC(N,J),CFAIL(N,J),(CUT(N,L),L=1,IQ)
71      FORMAT(13X,F4.1,4H TO ,F4.1,3X,10(4X,I2))
70      CONTINUE
          PRINT 62
          PRINT 72
230      *2 FORMAT(777/86H          (3) STATISTICS OF PRIOR AND POSTERIOR AT THE C
          PUTTING SCORE FOR GIVEN LOSS RATIOS///)
          PRINT 73,41(1,J),41(2,J),P,ME,PRMO,P,MSIG
73      FORMAT(38H          PRIOR/45H
          1          -----/32H          B
          2(,F5.2,1H,,F5.2,15H)          MEAN = ,F4.2/
          3,4
          4F4.2/56H          MODE = ,
          5,F4.2/77/          STAN DEV =
          PRINT 62
          PRINT 74,(CRIT(L,J),L=1,IQ)
235      74      FORMAT(63H          LOSS RATIO          POSTERIOR          CRITERION
          1 VALUES/43H          0=CSUCC TO CFAIL          STATISTIC ,
          2)(2X,F4.2))
          PRINT 62
240      DO 75 N=1,IR
          PRINT 76,CSUCC(N,J),CFAIL(N,J),(MEAN(N,L),L=1,IQ)
76      FORMAT(13X,F4.1,4H TO ,F4.1,19H          MEAN =,10(2X,F4.2))
          PRINT 77,(MODE(N,L),L=1,IQ)
77      FORMAT(43H          MODE =,10(2X,F4.2))
245      PRINT 78,(SIG(N,L),L=1,IQ)
78      FORMAT(43H          STAN DEV =,10(2X,F4.2))
75      CONTINUE
          PRINT 62
250      40 CONTINUE
          95 CONTINUE
          100 CONTINUE
          END

```

POSTERIOR PROBABILITIES(1) CUTTING SCORES(2) AND STATISTICS OF DISTRIBUTIONS(3) WHEN

TEST LENGTH = 5

PRIOR = $N(4.00, 1.00)$

GUESSING PARAMETER = .25

CARELESSNESS PARAMETER = .15

(1) TABLE OF $P_0(\theta | \text{GT CRITERION GIVEN SCORE})$
 = CONST TIMES INTEGRAL OF POSTERIOR FROM LIML TO LIMJ

SCORE	POSTERIOR	CONST	CRITERION VALUES					
			.70	.75	.80	.85	.90	.95
1	$N(5.33, 5.00)$	1.05	.05	.03	.01	.00	.00	.00
2	$N(6.00, 4.00)$	1.01	.17	.11	.06	.03	.01	.00
3	$N(7.00, 3.00)$	1.31	.40	.30	.21	.13	.06	.02
4	$N(8.00, 2.00)$	1.04	.69	.53	.48	.35	.22	.10
5	$N(9.00, 1.00)$	1.41	.90	.84	.76	.65	.50	.29
LIMITS OF THE			LIMJ =	.96	.95	.96	.96	.96
POSTERIOR INTEGRALS			LIML =	.75	.78	.82	.86	.89

(2) TABLE OF CUTTING SCORES

LOSS RATIO FAILURE TO FAIL	CRITERION VALUES					
	.70	.75	.80	.85	.90	.95
1.0 TO 1.0	3	4	4	4	5	5
2.0 TO 1.0	3	3	3	4	4	5
3.0 TO 1.0	2	2	3	3	4	4

(3) STATISTICS OF PRIOR AND POSTERIOR AT THE CUTTING SCORE FOR GIVEN LOSS RATIOS

PRIOR

(N 4.00, 1.00)

MEAN = .77

MODE = .85

STAN DEV = .15

LOSS RATIO FAILURE TO FAIL	POSTERIOR STATISTIC	CRITERION VALUES					
		.70	.75	.80	.85	.90	.95
1.0 TO 1.0	MEAN =	.70	.79	.79	.79	.87	.87
	MODE =	.75	.88	.88	.88	.86	.96
	STAN DEV =	.14	.12	.12	.12	.09	.09
2.0 TO 1.0	MEAN =	.70	.70	.70	.79	.79	.87
	MODE =	.75	.75	.75	.88	.86	.96
	STAN DEV =	.14	.14	.14	.12	.12	.09
3.0 TO 1.0	MEAN =	.63	.60	.70	.70	.79	.79
	MODE =	.63	.63	.75	.75	.88	.88
	STAN DEV =	.14	.14	.14	.14	.12	.12

BIBLIOGRAPHY

- ACER Development of the reading tests. Australian studies in school performance. Melbourne: Australian Council for Educational Research, 1976a.
- ACER Development of the numeration tests. Australian studies in school performance. Melbourne: Australian Council for Educational Research, 1976b.
- Alkin, M.C. 'Criterion-referenced measurement' and other such terms. In C.W. Harris, M.C. Alkin and W.J. Popham (Eds), Problems in criterion-referenced measurement. (CSE Monograph Series in Evaluation, No.3). Los Angeles: Centre for the Study of Evaluation, University of California, 1974.
- Besel, R. A mastery-learning test model. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1972.
- Besel, R. Using group performance to interpret responses to criterion-referenced tests (Professional Paper 25). Los Alamitos, California: Southwest Regional Laboratory for Educational Research and Development, 1973.
- Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability (Part 5). In F.M. Lord and M.R. Novick, Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley 1968.
- Block, J.H. Criterion-referenced measurements: potential. School Review, 1971, 69, 289-298.
- Block, J.H. Student learning and the setting of mastery performance standards. Educational Horizons, 1972, 50, 183-190.
- Block, J.H. Standards and criteria: A response. Journal of Educational Measurement, 1978, 15, 291-295.
- Bourke, S.F., and Koeves, J.P. (Eds). Australian studies in school performance Vol.3: The mastery of literacy and numeracy: First report (Education Research and Development Committee Report, No.13). Canberra: Australian Government Publishing Service, 1977.
- Box, G.E.P., and Tiao, G.C. Bayesian inference in statistical analysis. Reading, Massachusetts: Addison-Wesley, 1973.
- Burton, N.W. Societal standards. Journal of Educational Measurement, 1978, 15, 263-271.
- Cronbach, L.J. Response sets and test validity. Educational and Psychological Measurement, 1946, 6, 475-494.
- Cronbach, L.J. Further evidence on response sets and test design. Educational and Psychological Measurement, 1950, 10, 3-31.

- Cronbach, L.J. and Gleser, G.C. Psychological tests and personnel decisions (2nd ed.). Urbana, Illinois: University of Illinois Press, 1965.
- Davis, F.B., and Diamond, J.J. The preparation of criterion-referenced tests. In C.W. Harris, M.C. Alkin and W.J. Popham (Eds), Problems in criterion-referenced measurement. (CSE Monograph Series in Evaluation, No.3). Los Angeles: Centre for the Study of Evaluation, University of California, 1974.
- Davis, C.E., Hickman, J., and Novick, M.R. A primer on decision analysis for individually prescribed instruction. (ACT Technical Bulletin No.17). Iowa City, Iowa: The American College Testing Program. 1973.
- De Groot, M.H. Optimal statistical decisions. New York: McGraw-Hill, 1970.
- Diamond, J., and Evans, W.J. The correction for guessing. Review of Educational Research, 1973, 43, 181-191.
- Emrick, J.A. An evaluation model for mastery testing. Journal of Educational Measurement, 1971, 4, 321-326.
- Glaser, R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 519-521.
- Glaser, R., and Klaus, D.J. Proficiency measurement: Assessing human performance. In R.M. Gagné (Ed.), Psychological principles in systems development. New York: Holt, Rinehart, and Winston, 1962.
- Glaser, R., and Nitko, A.J. Measurement in learning and instruction. In R.L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington: American Council on Education, 1971.
- Glass, G.V. Standards and criteria. Journal of Educational Measurement, 1978, 15, 237-261.
- Hambleton, R.K. Testing and decision-making procedures for selected individualized instructional programs. Review of Educational Research, 1974, 44, 371-400.
- Hambleton, R.K. On the use of cut-off scores with criterion-referenced tests in instructional settings. Journal of Educational Measurement, 1978, 15, 277-290.
- Hambleton, R.K., and Cook, L.L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-96.
- Hambleton, R.K., and Novick, M.R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-169.

- Hambleton, R.K., Swaminathan, H., Algina, J., and Coulson, D.B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.
- Harris, C.W. Problems of objectives-based measurement. In C.W. Harris, M.C. Alkin and W.J. Popham (Eds). Problems in criterion-referenced measurement. (CSE Monograph Series in Evaluation, No.3). Los Angeles: Centre for the Study of Evaluation, University of California, 1974a.
- Harris, C.W. Some technical characteristics of mastery tests. In C.W. Harris, M.C. Alkin and W.J. Popham (Eds), Problems in criterion-referenced measurement. (CSE Monograph Series in Evaluation, No.3). Los Angeles: Centre for the Study of Evaluation, University of California, 1974b.
- Huynh, H. Statistical consideration of mastery scores. Psychometrika, 1976, 41, 65-78.
- International Mathematical Statistical Libraries, Inc. Sixth floor, GNB Building 7500 Bellaire Boulevard, Houston, Texas: Author, 1974.
- Kish, L. Survey sampling. New York: John Wiley, 1965.
- Lindley, D.V. Introduction to probability & statistics, Part 1: Probability. Cambridge: Cambridge University Press, 1965a.
- Lindley, D.V. Introduction to probability & statistics, Part 2: Inference. Cambridge: Cambridge University Press, 1965b.
- Livingston, S.A. Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 1972, 9, 13-26.
- Lord, F.M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.
- Lord, F.M., and Novick, M.R. Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley, 1968.
- Macready, G.B., and Dayton, M.C. The use of probabilistic models in the assessment of mastery. Journal of Educational Statistics, 1977, 2, 99-120.
- Macready, G.B., and Merwin, J.C. Homogeneity within item forms in domain referenced testing. Educational and Psychological Measurement, 1973, 33, 351-360.
- Meskauskas, J.A. Evaluation models for criterion-referenced testing: views regarding mastery and standard testing. Review of Educational Research, 1976, 46, 133-158.
- Millman, J. Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 1973, 43, 205-216.

- Millman, J. Criterion-referenced measurement. In W.J. Popham (Ed.), Evaluation in education: Current applications. Berkeley, California: McCutchan, 1974.
- Novick, M.R. Educational decisions with limited information. In W.E. Coffman (Ed.), Frontiers of educational measurement and information systems - 1973. Boston: Houghton Mifflin, 1973.
- Novick, M.R., and Jackson, P.H. Statistical methods for educational and psychological research. New York: McGraw-Hill, 1974.
- Novick, M.R., and Lewis, C. Prescribing test lengths for criterion-referenced measurement. In C.W. Harris, M.C. Alkin and W.J. Popham (Eds), Problems in criterion-referenced measurement. (CSE Monograph Series in Evaluation, No.3). Los Angeles: Centre for the Study of Evaluation, University of California, 1974.
- Novick, M.R., and Lewis, C., and Jackson, P.H. The estimation of proportions in m groups. Psychometrika, 1973, 38, 19-45.
- Pearson, K. (Ed.). Tables of the incomplete beta-function (2nd ed.). Cambridge: Cambridge University Press, 1968.
- Popham, W.J. Selecting objectives and generating test items for objectives-based tests. In C.W. Harris, M.C. Alkin and W.J. Popham (Eds), Problems in criterion-referenced measurement. (CSE Monograph Series in Evaluation, No.3). Los Angeles: Centre for the Study of Evaluation, University of California, 1974.
- Popham, W.J. As always, provocative. Journal of Educational Measurement, 1978, 15, 297-300.
- Popham, W.J. and Husek, T.R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.
- Pratt, J.W., Raiffa, H., and Schlaifer, R. Introduction to statistical decision theory. New York: McGraw-Hill, 1965.
- Raiffa, H., and Schlaifer, R. Applied statistical decision theory. Cambridge, Massachusetts: Harvard University Press, 1961.
- Roudabush, G.E. Models for a beginning theory of criterion-referenced tests. Paper presented at the meeting of the National Council for Measurement in Education, Chicago, April 1974.
- Shaycoft, M.F. A paradox in setting cutting scores on criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1976.
- Spinetti, J.P., and Hambleton, R.K. A computer simulation study of tailored testing strategies for objectives-based instructional programs. Educational and Psychological Measurement, 1977, 37, 139-158.

Sumner, R., and Robertson, T.S. Criterion referenced measurement and criterion referenced tests: Some published work reviewed. Slough, England: National Foundation for Educational Research in England and Wales, 1977.

Swaminathan, H., Hambleton, R.K., and Algina, J. A Bayesian decision-theoretic procedure for use with criterion-referenced tests. Journal of Educational Measurement, 1975, 12, 87-99.

Winkler, R.L. The assessment of prior distributions in Bayesian analysis. Journal of the American Statistical Association, 1967, 62, 777-800.



Occasional Paper 13

This paper reports a study which developed a criterion-referenced measurement model with corrections for guessing and carelessness within a Bayesian beta-binomial decision-theoretic framework. The results of a computer simulation of the model indicate that guessing and carelessness may markedly affect the determination of passing (cutting scores) for criterion-referenced tests, and hence the accuracy of decisions about mastery. The model has been applied in the calculation of cutting scores for tests of literacy and numeracy from the Australian Studies in School Performance project recently completed by the Australian Council for Educational Research

AUSTRALIAN COUNCIL FOR EDUCATIONAL RESEARCH

ISBN 0 85563 192 9

98

38 ARCO@W.A.U.